REPORT RESUMES

ED 012 891 AL 000 261

RESEARCH IN GERMAN-ENGLISH MECHANICAL TRANSLATION.

BY- LEHMANN, W.P. TOSH, L.W.
TEXAS UNIV., AUSTIN, LINGUISTICS RES. CTR.

REPORT NUMBER LRC-67-AFSC-4 PUB DATE APR 67
ROME AIR DEVELOPMENT CENTER, GRIFFISS AFB, N.Y.
REPORT NUMBER RADC-TR-67-98
EDRS PRICE MF-$0.50 HC-$4.48 112P.

DESCRIPTORS- *MACHINE TRANSLATION, *GERMAN, *ENGLISH, *PHRASE STRUCTURE, NOMINALS, LANGUAGE TYPOLOGY, MORPHOLOGY (LANGUAGES), TRANSFORMATIONS (LANGUAGE), SYNTAX, IBM 7040, LANGUAGE TRANSLATION SYSTEM,

UNDER CONTRACT WITH THE AIR FORCE, THE LINGUISTICS RESEARCH CENTER OF THE UNIVERSITY OF TEXAS CONDUCTED A RESEARCH PROJECT DESIGNED TO DEVELOP A GERMAN-ENGLISH SYNTACTIC TRANSLATION SYSTEM FOR SCIENTIFIC AND TECHNICAL TEXTS. MORE SPECIFICALLY, THE OBJECTIVES WERE TO (1) WRITE A GERMAN-ENGLISH TRANSFER GRAMMAR THAT WOULD LINK THE LINGUISTIC DESCRIPTIONS OF SOURCE AND TARGET LANGUAGES INVOLVED, AND (2) GIVE LINGUISTIC DESCRIPTIONS OF BOTH LANGUAGES THAT WOULD GENERATE MACHINE TRANSLATIONS OF GERMAN TEXTS. WITHIN A TRANSFORMATIONAL FRAMEWORK FOR INTERLINGUAL TRANSFER CODING, THE LANGUAGE TRANSLATION SYSTEM (LTS) ANALYZED TEXTS OF 500,000 GERMAN WORDS AND ONE MILLION ENGLISH WORDS. THE ANALYSIS YIELDED A DETAILED DESCRIPTION OF GERMAN SYNTAX WITH REGARD TO CERTAIN TYPES OF NOUN AND VERB PHRASES BOTH IN THE MONOLINGUAL AND INTERLINGUAL MODES. FOUR TRANSLATION RUNS WERE MADE, THE FIRST TWO ON A LEXICAL LEVEL AND THE LAST TWO ON MORPHOLOGY AND NP (NOUN PHRASE) STRUCTURES. (FB)

ED012891

RADC-TR-67-98
Final Report

RESEARCH IN GERMAN-ENGLISH MECHANICAL TRANSLATION

Dr. W. P. Lehmann
Dr. L. W. Tosh
The University of Texas

TECHNICAL REPORT NO. RADC-TR-67-98

April 1967

This document is subject to special export controls and each transmittal to foreign governments, foreign nationals or representatives thereto may be made only with prior approval of RADC (EMLI), GAFB, N.Y. 13440

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Rome Air Development Center
Research and Technology Division
Air Force Systems Command
Griffiss Air Force Base, New York

AL 000 261

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded, by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacturer, use, or sell any patented invention that may in any way be related thereto.

Do not return this copy. Retain or destroy.

RESEARCH IN GERMAN-ENGLISH MECHANICAL TRANSLATION

Dr. W. P. Lehmann
Dr. L. W. Tosh
The University of Texas

This document is subject to special export controls and each transmittal to foreign governments, foreign nationals or representatives thereto may be made only with prior approval of RADC (EMLI), GAFB, N.Y. 13440

"PERMISSION TO REPRODUCE THIS [redacted] MATERIAL HAS BEEN GRANTED
BY Z. Pankowicz,
RADC
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE [redacted] OWNER."

AFLC, GAFB, N.Y., 4 May 67-144

FOREWORD

This final report was prepared by The University of Texas, Linguistics Research Center, Box 7247, University Station, Austin, Texas, under Contract AF30(602)-3991, Project 4599, Contractor Report No. LRC 67 AFSC 4. The period of time covered was 1 January 1966 through 30 December 1966.

RADC Project Engineer was Zbigniew L. Pankowicz, EMIIH.

This document is not releasable to CFSTI because it contains information embargoes from release to Sino-Soviet Bloc Countries by AFR 400-10, "Strategic Trade Control Program."

This technical report has been reviewed and is approved.

Frank J Tomaini

Approved: FRANK J. TOMAINI
Chief, Information Processing Branch
Intel and Info Processing Division

James J. Dimel

Approved: JAMES J. DIMEL, Colonel, USAF
Chief, Intel and Info Prcs Div

FOR THE COMMANDER:
IRVING J. GABELMAN
Chief, Advanced Studies Group

ABSTRACT

Details are reported for syntactic description of German and English for use in the automatic Language Translation System of the Linguistics Research Center.

CONTENTS

Abstract iii

1 Introduction 1- 1

2 Summary of Research 2- 1

2.1 Objectives 2- 1
2.2 Translation 2- 2

3 German 3- 1

3.1 Syntax 3- 1

3.1.1 Problems in Grammar Design 3- 1
3.1.2 Noun Phrase 3-15

3.1.2.1 Adjective Phrase 3-25
3.1.2.1A German Pre-nominal Past Participles 3-29a
3.1.2.2 Post-nominal Attributes 3.30
3.1.2.3 Appositives 3-38
3.1.2.4 Conjunctions 3-46

3.1.3 Clause 3-52

3.1.3.1 Verb Phrase Rules 3-61
3.1.3.2 Prefixed Verbs 3-66

3.2 Lexicography 3-69

3.3 Test Corpora 3-74

4 English 4- 1

4.1 Syntax 4- 1

4.1.1 Noun Phrase 4- 1

4.1.1.1 Determiner Strings 4- 1
4.1.1.2 Adjective Strings With Connectors 4- 2
4.1.1.3 Adjective Strings 4- 2
W/out Connectors 4- 2
4.1.1.4 Post-nominal Modifiers 4- 3

4.1.2 Verb Phrase Description 4- 4
4.1.3 Clause 4- 4

4.1.3.1 Interrogative Clauses 4- 6
4.1.3.2 Subordinate Clauses 4- 6

4.2 Word Formation 4- 7

4.2.1 Lexicography 4- 7
4.2.2 Webster Morphology 4- 8

4.3 Test Corpora 4- 9
4.4 Concordances 4-10

5 Conclusion 5- 1

References R- 1

Personnel P- 1

Evaluation of RADC-TR-67-98, "Research in German-English Mechanical Translation", The University of Texas

1. Subject TR describes results of a 12-month R&D in German-English mechanical translation, started at the University of Texas in October 1959 and continuing since May 1959 under the sponsorship of the United States Army Signal Corps and the United States Army Electronic Laboratories.

2. The effort was directed toward the development of a German-English machine translation system capable of producing syntactic level translations of German scientific and technical texts. The effort had two distinct objectives: (a) writing a German-English transfer grammar that will link the linguistic descriptions of source and target languages involved, and (b) linguistic descriptions of both languages, including programming system, which will generate machine translation of German texts. Testing was restricted primarily to performing monolingual analysis of German and English.

3. Considerable progress was achieved in description of noun phrase, in both monolingual and interlingual mode, excluding relative clause. Some progress was also attained in description of verb phrase.

4. The effort was not directed toward developing an immediate, or even intermediate capability of translating randomly selected texts, and translations attempted within the framework of this effort were purely experimental.

ZBIGNIEW L. PANKOWICZ
Technical Evaluator

1 INTRODUCTION

In the following report we relate in some detail the descriptive linguistic effort pursued under Rome Air Development Center Contract No. AF 30(602)-3991 in support of the development of a German-English syntactic translation system. At the level of development reported here, we have made considerable progress in the description of the noun phrase, both in the monolingual and interlingual mode. Descriptions are comprehensive, excluding primarily the relative clause. Progress was made in the description of the verb phrase.

Testing of linguistic data during the year was limited largely to performing monolingual analysis in the Language Translation System (LTS). The size of either the German or English grammars has already grown to such proportion as to put significant strain on the processing capacity of the IBM 7040 system used in our research. As a result the most reliably interpretable output was obtained in the monolingual analysis mode. The translation runs performed, however, produced the results expected and were, therefore, considered satisfactory.

The descriptive effort in German is reported in detail. English research, which parallels the German, is reported more briefly. The German section, which follows the opening summary, begins with an analysis of some general problems of grammar design as executed in LTS. Specific details of German syntax are discussed, followed by a section on English and a concluding statement.

2 SUMMARY OF RESEARCH

2.1 Objectives

The general objective of work under the contract has been to develop further a syntactic translation capability within the framework of the Language Translation System (LTS). To this end we have designed and coded additional transfer grammar linking the two German and English monolingual syntactic descriptions. The grammars were verified in the processing system against appropriately selected samples of text data in each language from a stock of over 500,000 running words in German and 1,000,000 in English. Samples used for testing were drawn largely from the physical sciences. A brief survey is given below of the translation statistics developed and some of the problems encountered.

German and English grammar data compiled under earlier contracts [1] were expanded during the 12-month period. Grammatical descriptions are in context free phrase structure form with transformational facility provided in the structure of the interlingual transfer coding. The table below summarizes linguistic data statistics for the contract period.

	Grammar		Transfer	
	Dictionary	Syntax	Dictionary	Syntax
German	41,471	4,743	N.A.	N.A.
	43,123	7,518	13,139	1,109
English	100,263	7,718	N.A.	N.A.
	186,871	4,400	137,025	4,700

Figures are given in pairs indicating rule count, the upper member being the value for January, 1966, the lower for January, 1967. Statistics are not available for transfer data in January, 1966.

The increase in German dictionary entries is nominal, since we concentrated primarily on syntax acquisition. Corresponding transfer entries are based on some 10,000+ most frequent items in the general language. Syntax transfer coverage is discussed on detail below. Most of the increase in English dictionary and corresponding transfer is due to acquisition of data from the Russian Master Dictionary processed under Russian-English contract AF 33(657)-12950. The decrease in English syntax results from design changes in grammar allowing compaction.

2.2 Translation

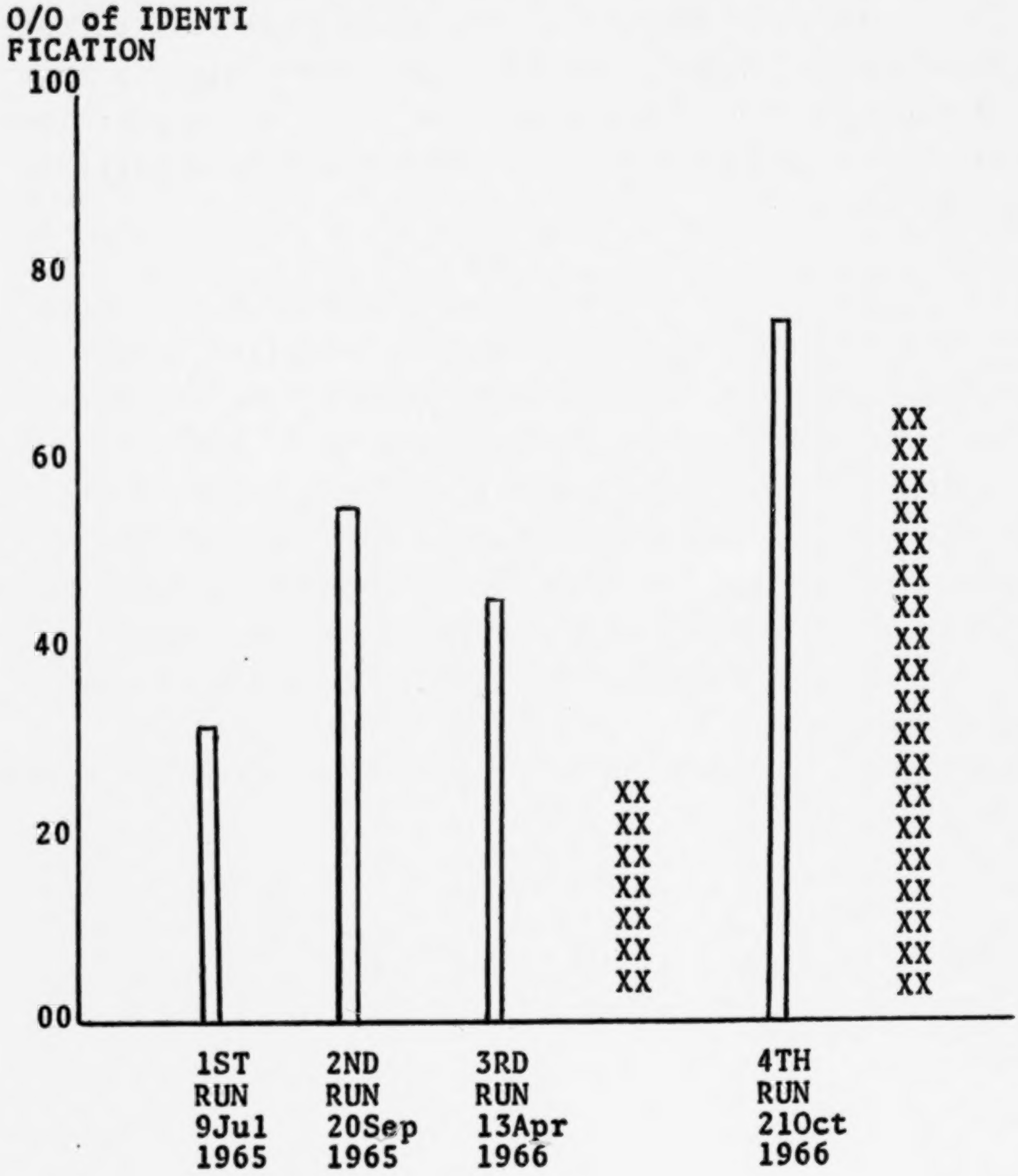
Ten paragraphs selected at random from a German text used for an earlier attempt were translated. In this translation run the optimized grammar was applied, i.e. only those grammar rules were available for which transfer rules had been written and compiled. The analysis was displayed and its results were evaluated along with the translation output.

(a) Analysis - German grammar and transfer rules were available for the following noun phrase patterns:

DET + ADJPH + NO
DET + ADJPH
DET + NO
ADJPH + NO

All noun phrases of the above type were correctly analyzed except those which contained items missing from the lexical part of our grammar and those which contained verbal forms, e.g. nominalized infinitives or adjectival participles.

(b) Translation - Four translation runs were made over a period of several months. English output improvement is shown in the graph below. The percentage of translated word stems is indicated by *11* and the percentage of translated words (including inflectional endings) by *XX*. The first two runs were made on the lexical level while the last two included morphology and (in the German grammar and transfer) the noun phrase structures given under (a). The choice of a different input text for the last two translations caused the percentage of recognition to drop somewhat from the second to third runs.


0/0 of IDENTI FICATION
100
80
60
40
20
00
1ST RUN 9Jul 1965
2ND RUN 20Sep 1965
3RD RUN 13Apr 1966
4TH RUN 21Oct 1966

3 GERMAN

A detailed study of the general problems of grammar design and its performance in the Language Translation System (LTS) spanned the 12-month contract period. The discussion below presents a detailed survey of the research in German syntax, lexicography, test corpora, and concordances.

3.1 Syntax

3.1.1 Problems in Grammar Design

The problem can be summarized thus: Guarantee an optimal translation with a minimum of distinctions in the source grammar, making optimal use of the LTS software.

The power of the optimized grammar is dependent on the number of constants dominated by a transfer (TRN)-rule, i.e. there is no optimization when all TRN-rules have only one term on their right. There is the "more optimization" when more terms are on the right of TRN-rules. This seems to favor low-degree syntactic rules.

The application of TRN-rules is dependent on the probability of the underlying syntactic strings. If a sequence (A+B) + (C+D) is analyzed by three syntactic rules of degree D2, a D4 TRN-rule covering this will have the same probability as a combination of 3 D2 TRN-rules. The underlying syntactic tree is the same in both cases.

If two different syntactic analyses derive the same (non-terminal) string, one of them assigning wrong superscripts, then both TRN-correspondences will be applied. In these examples a syntactic D4 rule derived the same string as a sequence of three syntactic D2 rules. Because of probability of the syntactic D4 rule it would have been preferred. The others would, however, not be rejected and result in an incorrect alternative translation.

It thus seems advisable to prevent improper application of established syntactic rules by increasing the number of clause-level elements indicating positional values and environment, thus restricting or preventing certain concatenations. This would not increase the number of clause rules.

Generalizing D1 rules had two disadvantages: They allowed re-distribution of probabilities, e.g. if A dominates B(1), B(2), ... B(5) only, each rule has a probability of 1/5. If we introduce generalizing rules A dominating G(1), G(2) where G(1) dominates B(1-3) and G(2) dominates B(4,5), the sequence A-B(1-3) has a probability of 1/6 each, A-B(4,5) one of 1/4 each. This can be overcome by weighting the rules A-G(1,2). The given weight is the number of rules with G(X) on the left side. Thus A-G(1) = W3 and A-G(2) = W2, the probability of A-G(1) = 3/(3+2) = 3/5, of A-G(2) = 2/(3+2) = 2/5. A-B(1-3) has a probability of 3/5 X 1/3 = 1/5, A-B(4,5) one of 2/5 X 1/2 = 1/5, for each sequence A-B(X). (Probability of a rule: the weight of the rule with X as first term divided by the sum of the weights of all rules with X as their first term.)

The main difficulty is, however, the increase of co-terminating rules which cannot be affected by the above weighting method. (Co-terminating rules: the set of rules or rule sequences analyzing all the text-intervals ending in one specified position. They may begin at the specified position or anywhere left of it). A table containing the probabilities of the corresponding co-terminating rules is constructed for each position in the sample. Each table can maximally hold 64 probabilities. All rules or rule sequences with the same probability have one common entry in the table. If the number of probabilities exceeds the specified maximum of 64, the lowest will be dropped. The rules or rule sequences corresponding to them will no longer be used for the given intervals. Thus in the above example the sequence A-B(1) without generalizing steps would make one entry in the table with a probability of 1/5. The sequence A-G(1)-B(1) would make two entries with probabilities of 1/3 (G(1)-B(1)) and 1/5 (A-G(1)-B(1)).

In the optimized grammar, tables are built only for rules which occur in TRN-rules. Each transfer rule causes, at most, one entry in the table. There may be other TRN-rules with the same probability. The number of constants in a TRN-rule is irrelevant for the number of table entries since they are used only to compute the probability of the entry. Generalizing steps, if they do not occur as the only constant in a TRN-rule, will not affect the analysis or translation. The problem reduces to a choice of simplifying syntax at the cost of complicating transfer, or vice versa. The number of TRN-rules would not, however, increase. "Complicating transfer" means to increase the number of terms within a TRN-rule.

The sentence *Edelmetalle treten hier auf* receives interpretations as CLS or SNTC over the following spans:

Co-terminating in *treten* :

alle treten
Metalle treten
Edelmetalle treten

Co-terminating in "*hier*":

alle treten hier   *Metalle*, *Edel-*, etc.

Co-terminating in "*auf*":

alle treten hier auf etc.

Co-terminating in " . "

alle treten hier auf. etc.

The inability to recognize word-boundary thus results in an increase of table entries not only for the position occupied by the final *e* of *Edelmetalle*, but also for the table of each subsequent element if this concatenates with the preceding elements (as in above examples). An indication of "beginning of word" by encoding an initial blank on preterminal level would reduce the number of table entries depending on *Edelmetalle* by two-thirds. With longer German sentences the reduction, especially at the end of a sentence, would be even greater. This problem is not affected by the optimized grammar. These partial sentences are well-formed and corresponding TRN-rules will exist.

Certain syntactic morphemes depend on lexical items, e.g. gender, case, etc., and cannot be directly translated into another language because it may not have them or employs them differently according to the lexical entry on which they depend. Such morphemes need only be indicated as subscripts when they are necessary for translation.

Thus it was decided to drop gender distinction in NP's in order to show agreement between noun and remote relative clause. It will remain difficult enough to describe this agreement without gender indication, e.g. *ein brot ... das* in corpus 68, sample 5:

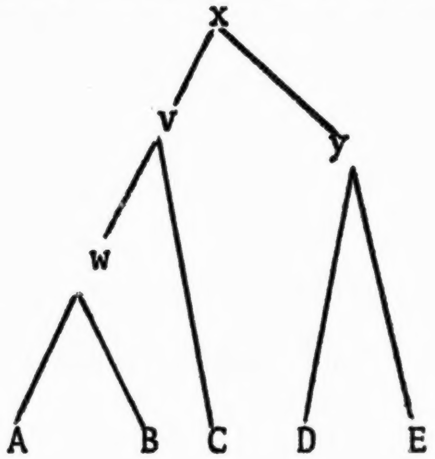
> *so wird am Ende der Konsument ein Brot verlangen, und die Fabriken werden es ihm liefern, wie es der bedeutende englische Ernaehrungsforscher Sir Edward Mellamby beschrieben hat, das ...*

The number of clause rules would be reduced by suppressing person-agreement between subject and predicate on clause-level as this can be provided in English through transfer. To suppress number agreement would result in further reduction. Since most German subjects and objects have the same form (except for masculine nominative and accusative and conjoined phrases containing them) we will get two interpretations with subject and accusative object. They will differ in their superscript assignments only. To include the rule predicate = predicate/number in TRN-clause rules will suffice only when object and subject have different numbers. We therefore need to include the rule subject = subject/number in the TRN-subtree.

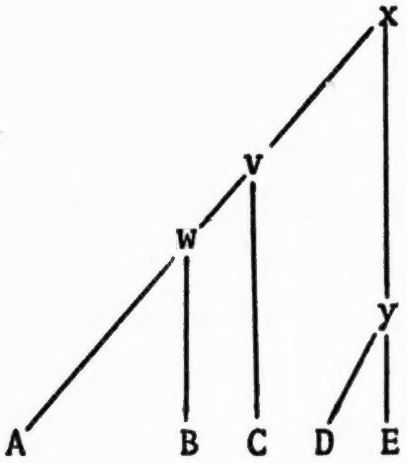
After the German-English semantic demonstration held February 16, 1966, under Contract DA 36-039 AMC 02162(E) the German syntactic clause-rules were preserved since they were not affected by the de-generalizing steps. Thus German syntax should perform rather well after the rules connecting clauses with the degeneralized noun-phrases and verbs are written.

Some difficulties not affected by de-generalizing were avoided or lessened by some modifications which will be developed discussing the following example.

When drawing tree diagrams we used graphs of this form:



This graph will now be changed to:

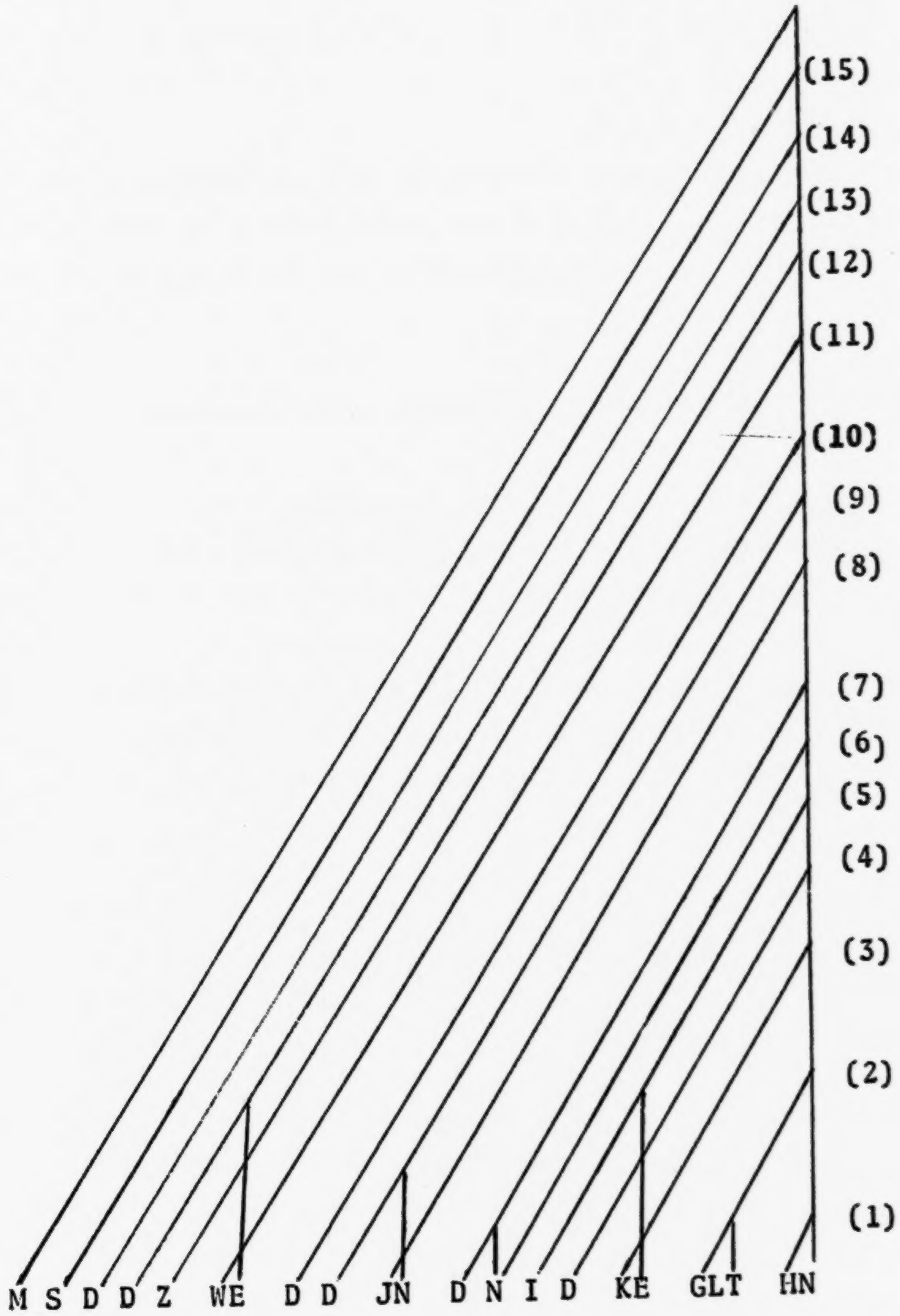


This form of diagram has the advantage of representing the analysis probability tables constructed by the analysis program for each character in a string of input text.

In the following example, the sentence

Man sah dort die zwei Woelfe,
die den Jungen die Nahrung in die
Koerbe gelegt hatten.

is relatively simple in structure and is representative of difficulties encountered in the German clause. A simplified parsing can be represented by:


(15)
(14)
(13)
(12)
(11)
(10)
(9)
(8)
(7)
(6)
(5)
(4)
(3)
(2)
(1)
M S D D Z WE D D JN D N I D KE GLT HN

Assuming that the rule and rule sequences analyzing spans left of and including the *N* of *HN* (*hatten*) have different probabilities following rule probabilities will be entered for *N*:

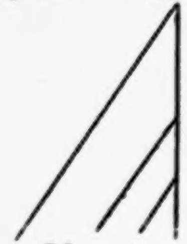
```
(1)    VVAX/P VVHAT    *EN          =1

(2)    VVB/PFLVVPAPL   VVAX/P       =1

(3)    VVB/PFLVOBJ/A   VVB/PFL      =4  (ADV=OBJ/A)
              VOBJ/G
              VSUBJ/P
              VADV

(4)    VVB/PFLVOBJ/A   VVB/PFL      =4  (As 3 without OBJ/G)

(5)    VVB/PFLVADV     VVB/PFL      =1

(6)    VVB/PFLVOBJ/A   VVB/PFL      =5  (As 3 plus OBJ/D)

(7)    VVB/FPFVOBJ/A   VVB/PFL      =3  (As 3 without OBJ/G)

(8)    VVB/PFLVOBJ/A   VVB/PFL      =5  (As 6)

(9)    VVB/PFLVOBJ/A   VVB/PFL      =3  (As 6 without SUBJ
                                          and OBJ/G)
(10)   VVB/PFLVOBJ/A   VVB/PFL      =4
       VVB/PFLVADV     VVB/PFL
       VCLS/RLVSBJ/RL  VVB/PFL
       VCLS/SBVSUBJ    VVB/PFL

(11)   VNO/GK VNO/GK   VCLS/RL      =1

(12)   VSUBJ  VNO/GK                =6  (For the interpretations
       VCMPL                                of Woelfe)
       VOBJ/S
       VOBJ/A
       VOBJ/G
       VADV   VOBJ/A
```

```
        VNP/PNAVNUMBERVNO/GK          =6
        VSUBJ  VNP/PNA
        VOBJ/S
        VOBJ/A
        VCMPL
        VADV

(13)    VNP/PNAVDET +VNUMBER+VNO/GK =6   (As above)
        VVB/S  VVB/S  VSUBJ          =3
                       VOBJ/A
                       VADV

(14)    VVB/S  VVB/S  VADV           =1

(15)    VCLS   VSUBJ  VVB/3          =1

                  PRELIMINARY TOTAL =55
```

In the above examples the expansion in points (1) - (9) had been performed for the element VVB/PFL, i.e. VPAPL+VAX. Since the German VPAPL can also be expanded and then concatenate with the final auxiliary we get additional probability entries when VPAPL has been expanded just once, just twice, etc. Finally in the above examples it was assumed that each element was uniquely concatenated with the final element. However, because of rules like VADV→VADV + VADV, the concatenations of the adverbial interpretations of *die*, *den*, *Jungen*, *die*, *Nahrung* which are all OBJ/A's and thus ADV's will concatenate with VPAPL and VB/PFL in groups of two, three, etc. Since comma conjoins NP's or OBJ's, the sequence beginning with *dort* to *Koerbe* will concatenate with VPAPL and VB/PFL. The number of probabilities entered for the *N* of *hatten* will be some hundreds, all of which will be dropped except the 64 highest.

Interpreting the above results, it seems that analysis difficulties result from the following:

(A) The shape of the German trees (left expanding)



(B) Duplication of Interpretations (above, first PAPL+AUX then expansion, and first expansion of PAPL, then concatenation with AUX).

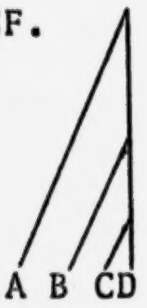
(C) Multiple concatenation of NP's as adverbs

(D) Multiple interpretation of elements on CLS-level

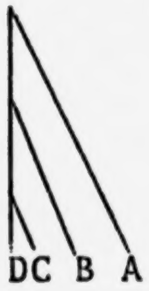
(E) Concatenations at very low level.

It might be interesting to point out that the above example is only minimally affected by introducing verb-object agreement, and not at all by the optimized grammar since all interpretations are well-formed and will occur in some contexts. Since the main difficulty with analysis is mainly the left-expanding structure of German, it may be interesting to point out that by mirroring German and its grammar these difficulties could be overcome without changes in linguistic description

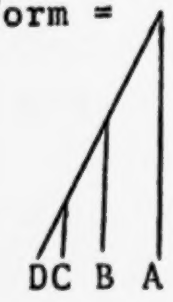
(CF.



Mirrored =



In program form =



)

The probabilities would now be distributed over the whole text instead of being massed at one point.

Since the German CLS-rules (main I and II, subordinate, relative, interrogative-hypothetic) were not affected by degeneralizing, only the rules connecting CLS-level with NP's and VPH's had to be written. Rules for *zu*-infinitive and *um-zu*-infinitive constructs were encoded. Modifications were made in the following areas:

Verb government is not a property of the finite verb alone but also of the non-finite verb, i.e. the verb stem. The pattern similarity, i.e. applying rule structures established in main clauses to such subordinate clauses, was given up. Thus prefix-verb concatenations are now analyzed as ((prefix+VBSTEM) +VB-ending). The old method of analysis would have resulted in maximally 27 different TRN-rules for each prefixed German verb. The inflectional ending is covertly contained in the TRN-rule; there are maximally 27 endings for a given German VB-class. The new solution allows us to regard PFX+VB as a degree zero TRN-entry which does not contain information as to inflections.

The rules are

```
VVCLASSVPRFX  VVCLASS
                 B
```

For the case that the verb-class did not immediately follow the prefix as with *ge*-past participles, e.g. *gelaufen, gelegt, abgelaufen, abgelegt,* two solutions were possible:

```
(1)  A 4-branch rule   VVPAPL VPRFX  *GE  VVCLASS*ending
                                      B

(2)  A 2-rule sequence VVPAPL VVPAPL *ending
                                GE
          and          VVPAPL VPRFX  VVPAPL
                         GE            GE
                                      B
```

We selected (2) to preserve symmetry with the prefix-verb (no ending included) and the following solution for *zu*-infinitives.

Of two solutions

```
        VVINF-ZVPRFX  VZU     VVCLASS      (abzuleg-en)

and     VVINF-ZVPRFX  VVINF-Z     and VVINF-ZVZU     VVCLASS
         STEM          STEM            STEM          B
                       -P              -P
                      B
```

we selected the latter. Whatever solution is finally used, we propose to give *zu* superscript 0, 2 or 3. This method will prevent most often the generation of the split infinitive in English. The TRN-rule of the prefix-*zu*-infinitive structure will be of degree one, excluding *zu*. This treatment of prefixed verbs is of some practical advantage, since the sequence prefix+VB occurs predominantly, the only exceptions are present and past active finites in main and interrogative clauses.

Verb structure and expansion rules are restricted so that at the lowest level there is a non-expandable verb-element to prevent unrestricted concatenation from left to right.

Every verb can be modified by an adverb. By writing a rule as

```
        VVERB    VADV    VVERB
         +ADV             NON-EX
```

which is non-recursive, all adverbs left of

VERB
NON-EX

are concatenated first before they concatenate with the verb-element, thus creating a somewhat tagmemic structuring of the adverbials in immediate environment.

Verb/NON-EX and Verb/+ADV are derived from expandable elements as VPAPL/IST, VPAPL/WRD, VPAPL/HAT, VINF/EXP, VB/EXP.

The first two can only dominate a dative object. The rules are thus non-recursive.

```
VVPAPL  VOBJ   VVPAPL
 X       D
```

or

```
VVPAPL  VOBJ   VVPAPL
 X       D      +ADV
```

to allow for subsequent expansion by ADV.

The last three VB-types are recursive for OBJ/D expansion:

```
VVTYPE/EXP    VOBJ/D  VVTYPE/EXP
```

They are non-recursive for OBJ/A expansion:

```
VVTYPE/EXP    VOBJ/A  VVERB
```

or

```
VVTYPE/EXP    VOBJ/A  VVERB
                        +ADV
```

Thus unique and, more important, correct superscripts are assigned to sequences as OBJ/D + OBJ/A + ADV + ADV + VERB. Multi-branch non-recursive rules were written for OBJ/A + OBJ/D sequences.

VV-TYPE/EXP

VOBJ/A	VOBJ/D	VVERB
S2	S3	S1

Or

VOBJ/A	VOBJ/D	VVERB +ADV
S2	S3	S1

and ADV + OBJ + VB's (analogous treatment). A similar solution was encoded for verbs that expand to the right.

Some distinctions like CMPL/SING and CMPL/PLUR were dropped because of examples like *die Roemer waren ein tapferes Volk.*

3.1.2 Noun Phrase

This section summarizes research completed in noun phrase description during January - July 1966. In order to avoid superfluous multiple analyses, it was decided to exclude the generalizing steps from the grammatical description, i.e. one-branch rules of the form

VNO	VNO
MN	MNDA PNGA

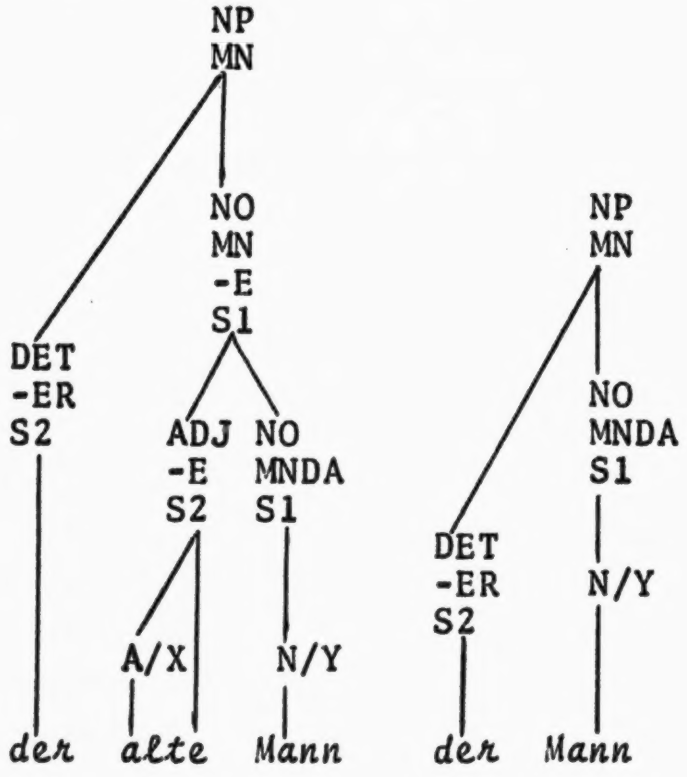
Or

VNP	VNO
X	X

Only one-branch rules coming down from the labels OBJ, SUBJ, ADV and ATTR were kept because elimination of such rules would make the number of necessary grammar rules unwieldy.

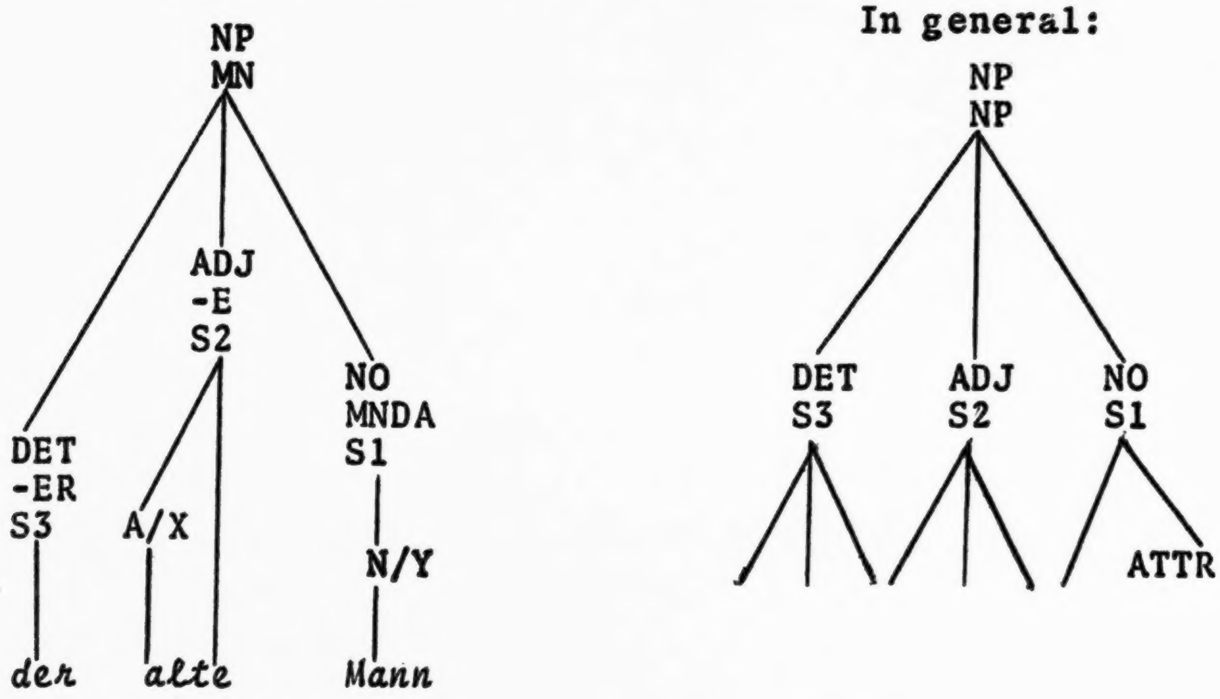
Taking these changes of the noun phrase description into consideration, two patterns of noun phrase analysis were possible: (1) binary and (2) non-binary.

(1) Binary description:


NP
MN
NO
MN
-E
S1
DET
-ER
S2
ADJ
-E
S2
NO
MNDA
S1
A/X
N/Y
der
alte
Mann
NP
MN
NO
MNDA
S1
DET
-ER
S2
N/Y
der
Mann

This description would require two different sets of nominal labels: those of unexpanded nominals, i.e. the full set of labels dominating the noun inflection steps and, secondly, those of nominals expanded by one or more preceding adjectives. The latter set of nominal labels would include the inflection of the preceding adjective as subscript. From this label (i.e. the combination of gender, case and adjective inflection) it can be seen whether a nominal includes a strong or a weak adjective. DET + NO concatenation rules could then be coded for the relevant nominals labels, i.e. unexpanded nominals and those including strong adjective information.

(2) Non-binary description:



The non-binary description was chosen for several reasons:

1. Ambiguities are resolved earlier since the determiner, which carries disambiguating information, is concatenated earlier with the rest of the noun phrase.

2. It is simpler and clearer because only one set of nominal labels is necessary (the labels dominating the noun inflecting rules).

Based on this decision, the following sets of rules were coded:

(a) Unmodified nominals

VSUBJ	VNO	(R1)
Number	Gender	
	Case	

In this and the following examples, (R1, R2, etc. denote rule number for cross reference.

VOBJ	VNO	(R1)
Case	Gender	
	Case	

VCMPL	VNO	
	Gender	
	Case	

VSUBJ	VPRN	(R3)
Number	Gender	
	Case	

VOBJ	VPRN	(R3)
Case	Gender	
	Case	

VSUBJ	VADJPH	(R2)
Number	Inflection	

```
VOBJ    VADJPH     (R2)
 Case    Inflec
         tion
```

As shown, this includes pronouns and nominalized adjectives. Examples are:

in der Luft ist Stoermaterial genug

sie beugen das Licht ab

kurzwelliges (Licht) wird noch in das Berliner Auge abgelenkt

Transfer over these rules was designed as follows:

```
VNO     CR1
D1

VNO     CR2
D1

VPRN    CR3
D1
```

(b) Noun phrases consisting of nominals modified by adjectives:

```
VNP      VADJPH    VNO            (R4)
 Gender   Inflec    Gender
 Case     tion      Case
         S2        S1
```

Example:

(in) prinzipieller Hinsicht

Transfer was coded

```
VNO+AJ CR4
D2
```

(c) Noun phrases consisting of determiner and nominal:

```
VNP       VDET     VNO              (R5)
 Gender   Inflec   Gender
 Case      tion    Case
          S2       S1

VNP       VDET     VADJPH           (R6)
 Gender   Inflec   Inflec
 Case      tion     tion
          S2       S1

VNP       *M       VNO              (R7)
 Gender             Gender
 Case               Case
 -M
```

Examples:

die Streuwirkung

das Fruchtlose ihrer Bemuehungen

(zu)r Zeit

The last of the above sets of rules was written for the partial determiners *r*, *s*, and *m* concatenated with the relevant nominals and adjectives. As indicated, these partial determiners are not classified but rather treated as constants.

Transfer:

```
VNO+DT CR5
D2

VNO+DT CR6
D2

VTHE+SGCR7
D1
```

(d) Noun phrases consisting of a nominal, an adjective phrase and a determiner:

```
VNP      VDET     VADJPH  VNO     (R8)
 Gender   Inflec   Inflec  Gender
 Case     tion     tion    Case
         S3       S2      S1

VNP      *M       VADJPH  VNO     (R9)
 Gender            Inflec  Gender
 Case              tion    Case
 -M               S2      S1
```

Examples:

das spezielle Relativitaetsprinzip

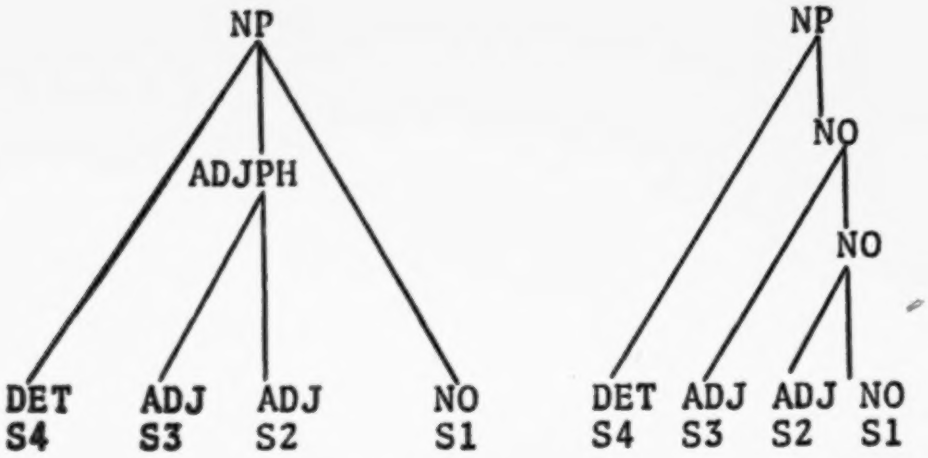
(zu)*m naemlichen Zeitpunkt*

Transfer over these sets of grammar rules was coded

VNO+AJ+CR8
 DT
D3

VTHE+SGCR9
 +AJ
D2

Strings of the form DET + ADJ + ADJ + NO have topologically different descriptions in German and English.


German

English

NP
ADJPH
DET
S4
ADJ
S3
ADJ
S2
NO
S1
NP
NO
NO
DET
S4
ADJ
S3
ADJ
S2
NO
S1

Therefore, transfer had to be written over the whole string. The full set of these transfer rules was coded. According to conventions, the label is VNO+AJ+ (D4).
AJ+DT

Transfer was coded for the grammar rules concatenating all nominals with ATTR/B. These rules have the form

VNO+AJ   C...
D2

For nominalized adjectives, which are frequently represented by NO in English, we coded transfer over that set of rules which concatenates DET + ADJPH under a NP label. These rules are in the same transfer class as the corresponding transfer rules over DET + NO:

```
VNO+DT   C...
```

The necessary rules for transfer of uninflected adjectives as adverbs were coded:

```
V-LY      CR1      CR2
D1
```

where R1 is the rule number of

```
VADV      VAV
```

and R2 the set of grammar rules like

```
VAV       VAI/O+S              etc.
```

Adjective inflection rules were included in the TRN classes SG and PL respectively to allow translation of German nominalized adjectives into English nouns.

Analogous to the description of nouns, rules of the form

```
VPRN/DMVDET
 Gender Inflec
 Case   tion
```

were changed to preserve the genuine ambiguity in respect to gender and case. Transfer was adjusted accordingly.

Example:

diesem kann man dadurch entgehen, dass...

It was decided not to carry the ambiguity in the pronoun label up to the clause level elements (OBJ, SUBJ) since this would increase the number of necessary clause rules to unreasonable size.

The description of pre-posed adverbs which modify a whole noun phrase was left unchanged but was completed to include all new noun phrase and pronoun labels which resulted from the elimination of generalizing steps. These rules have the form

```
VNP        VAV       VNP
 Gender     PRE-NP    Gender
 Case      S2         Case
                     S1

VPRN/DMVAV           VPRN/DM
 Gender PRE-NP        Gender
 Case   S2            Case
                     S1
```

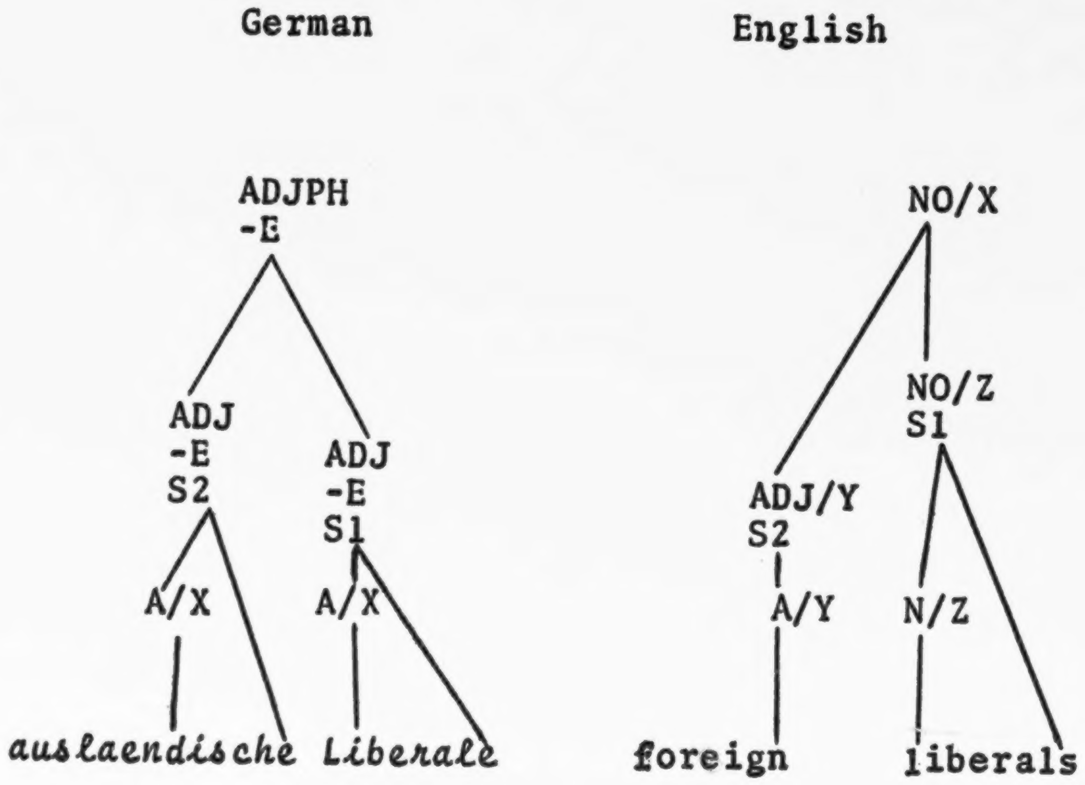
Examples:

sogar das kristalline Quarz

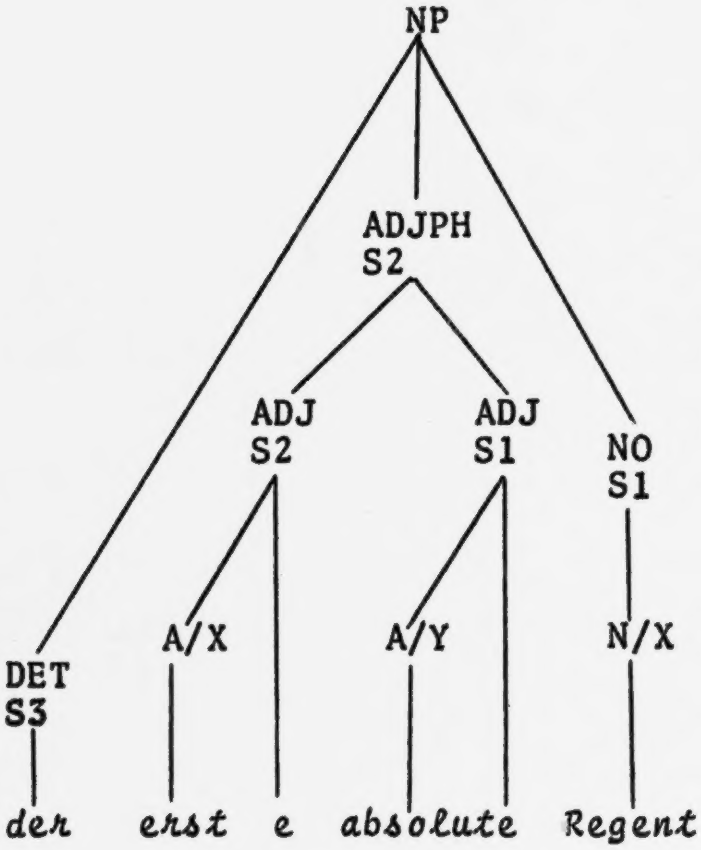
sogar diese

3.1.2.1 Adjective Phrase

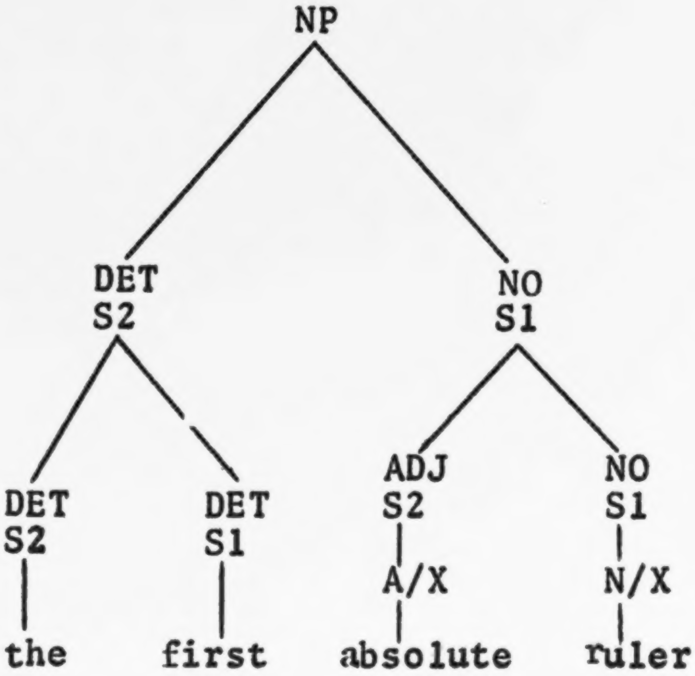
To guarantee superscript correspondence between English and German even in those cases where an item is analyzed as noun in English and as adjective in German or those where the translation of a German adjective is an English determiner, the superscripts for the above rules were assigned as shown, i.e. conjunctions were given the lowest superscript(s) in left-to-right order, adjectives were given the next higher superscripts in right-to-left order.


German
English
ADJPH
-E
ADJ
-E
S2
ADJ
-E
S1
A/X
A/X
auslaendische
Liberale
NO/X
NO/Z
S1
ADJ/Y
S2
A/Y
N/Z
foreign
liberals

German


NP
ADJPH
S2
ADJ
S2
ADJ
S1
NO
S1
A/X
A/Y
N/X
DET
S3
der
erst
e
absolute
Regent

English


NP
DET
S2
NO
S1
DET
S2
DET
S1
ADJ
S2
NO
S1
A/X
N/X
the
first
absolute
ruler

Adverb-adjective concatenation rules were written recursively under the adjective node with the adjective carrying the lowest superscript since it is considered the phrase head.

The present German grammar contained two sets of rules concatenating adverbs with adjectives.

VADJPH	VADV	VADJPH
-X	S2	-X
		S1

For examples like

der ohnehin schon intensive und
bittere Kampf,

and

VADJ	VADV	VADJPH
-X	S2	-X
		S1

For examples like

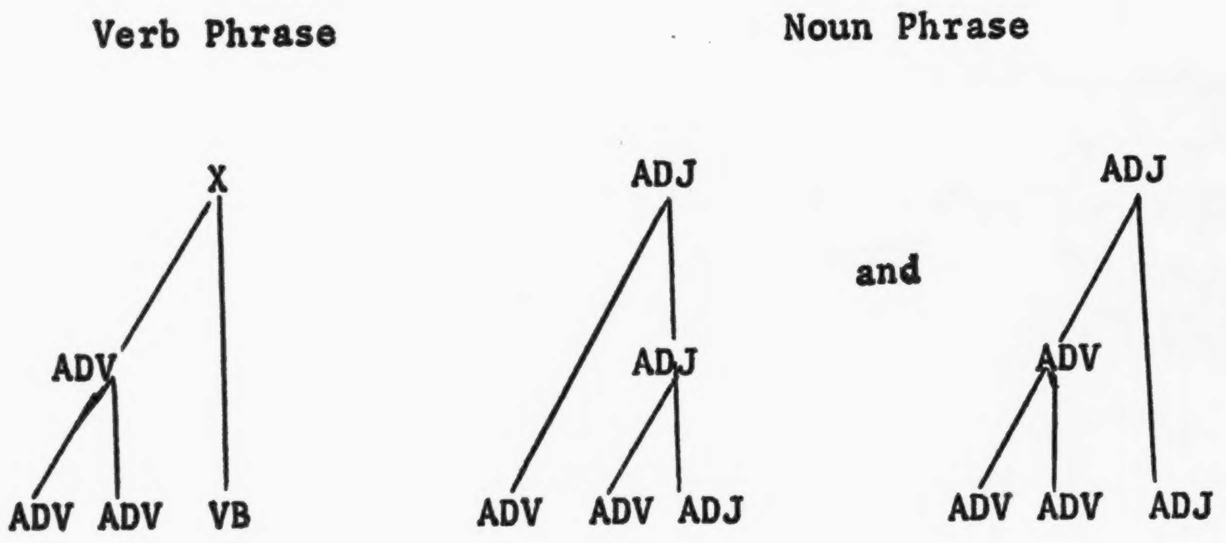
der wichtige und gut verdeckte
Abdruck.

Both sets of rules will be applicable regardless of whether the adverb actually modifies both adjectives or only the first one.

Since one of the two sets of rules would suffice for analysis and translation and, mostly, to avoid the double analysis for every string of ADV + ADJ, Set 1 was eliminated.

A distinction between adverbs occurring on clause level and in verb phrases on the one hand and those occurring in noun phrases on the other hand was built into the grammar for the following reason:

ADV + ADV concatenation is necessary for verb phrase description where ADV + VB are not concatenated recursively under the VB label. Multiple analyses within the noun phrase result where ADV + ADJ are concatenated recursively under the adjective label.


Verb Phrase
Noun Phrase
X
ADV
ADV ADV VB
ADJ
ADJ
ADV ADV ADJ
and
ADJ
ADV
ADV ADV ADJ

To avoid the multiple analyses shown above, the new description of adverb sequences within the noun phrase is as follows:

The only adverb label used within the noun phrase is AV which is defined in the grammar by rules of the form

VAV	*OFT, HEUTE ETC.		(Dictionary class)
VAV	VA CLASS		(*schnell, schoen* etc.)
VAV	VPRPH		(*am Anfang*)
VAV	VAV	VCONJ/BVAV B	(*hier und dort*)

For use in verb phrases and on clause level the following rule exists:

```
VADV  VAV
```

but only AV concatenates with adjectives:

```
VADJ  VAV    VADJ
```

Since adjectives are not subclassified according to possible case government (a feature necessary for synthesis only), dative and genitive objects were concatenated with ADJPH lables rather than terminal adjective classes.

A total of 139 rules was coded for transfer of adjectives in the comparative and superlative. These rules have the form:

```
VCMPRIVCR1    CR2    CR3
D1

VSPRLTVCR1    CR2    CR3
D1
```

where R1 is a GRM rule of the type

```
     VADJPH  VADJ
      -X      -X

R2:VADJ    VAZ   *X
    -X           B

R3:VAZ     VAY   *Comparative or superlative inflection.
                 B
```

3.1.2.1A German Pre-nominal Past Participles

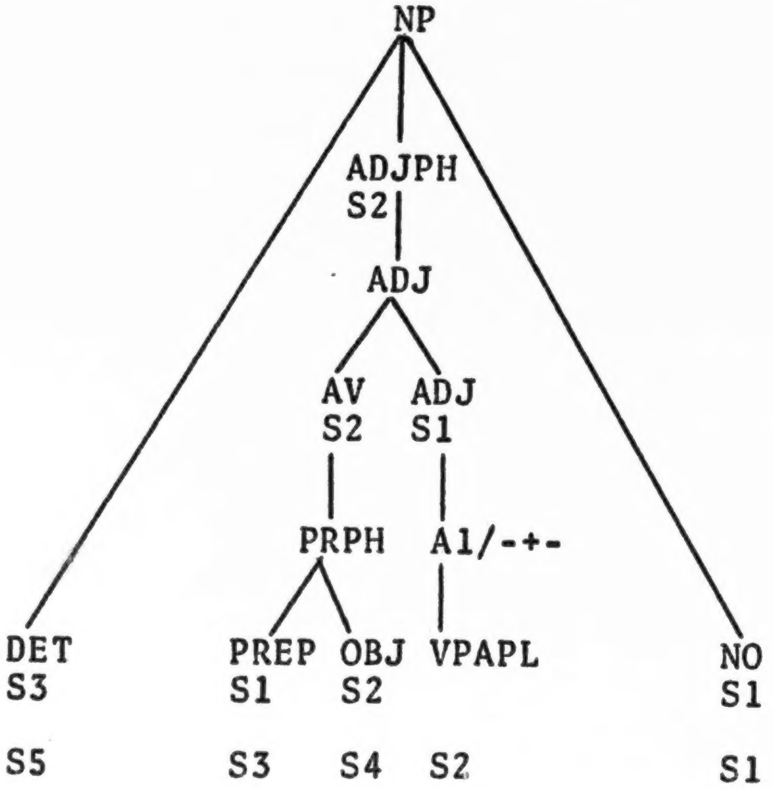
Past Participles Modified by Prepositional Phrases

This structure is the most frequent among past participial phrases in the German physical science concordance and is translated into post-nominal participial phrases in English.

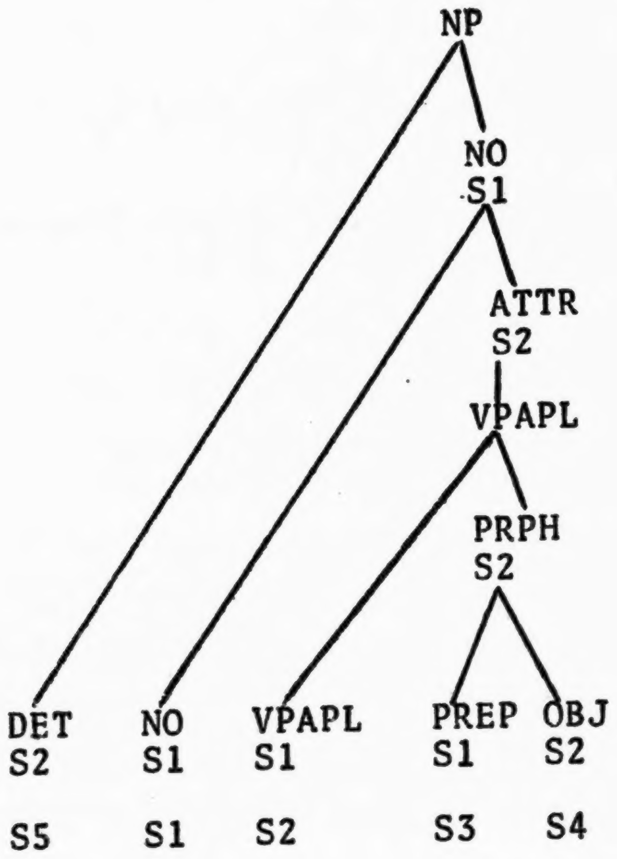
One past participle modified by a prepositional phrase:

German *den von Photonenraketen angetriebenen Raumfahrzeugen*
English *the space craft propelled by photon rockets*

German


NP
ADJPH
S2
ADJ
AV
S2
ADJ
S1
PRPH
A1/-+-
DET
S3
PREP
S1
OBJ
S2
VPAPL
NO
S1
S5
S3
S4
S2
S1

English


NP
NO
S1
ATTR
S2
VPAPL
PRPH
S2
DET
S2
NO
S1
VPAPL
S1
PREP
S1
OBJ
S2
S5
S1
S2
S3
S4

The past participle can be modified by both a prepositional phrase and a normal adverb. In this case, too, the past participle with all its modifiers is in post-nominal position in English, but the adverb can usually precede the past participle while the prepositional phrase, as usual, follows it. Examples:

German *der gleichfalls an Bord gemessenen Zeit*
English *of the time also measured on board*

German *der von Fermat ausdruechklich behaupteten Theoreme*
English *of the theorems expressly stated by Fermat*

German *ein von uns willkuerlich eingefuerhrtes Koordinatensystem*
English *a coordinating system randomly introduced by us*

Two Past Participles Modified by a Prepositional Phrase

Both past participles can be translated into postnominal structures in English, the prepositional phrase following the past participle. Example:

German *ein mit Glaswolle beschicktes und auf 150 Grad C erhitztes Rohr*
English *a pipe filled with glass-wool and heated to 150 degrees C.*

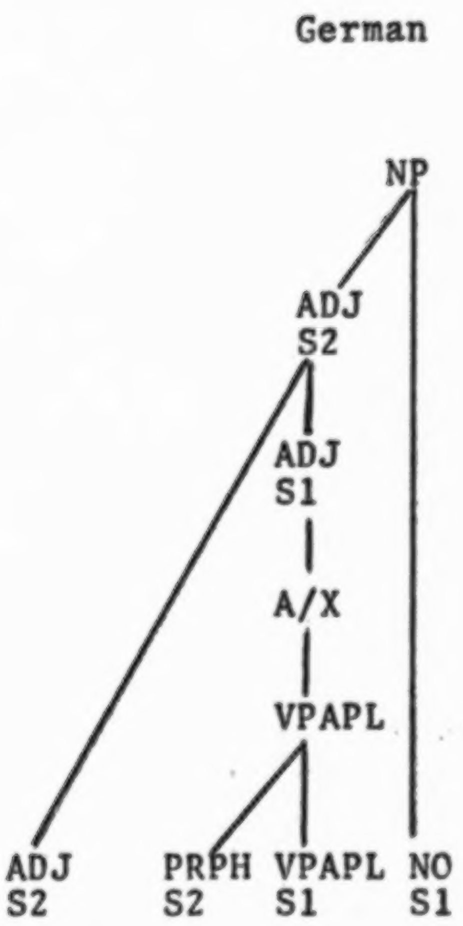
One Past Participle Modified by Prepositional Phrase and One or More Adjectives

In these cases the adjective(s) precede(s) and the past participle follows the nominal in English, no matter what the order of past participle and adjective is in German. Examples:

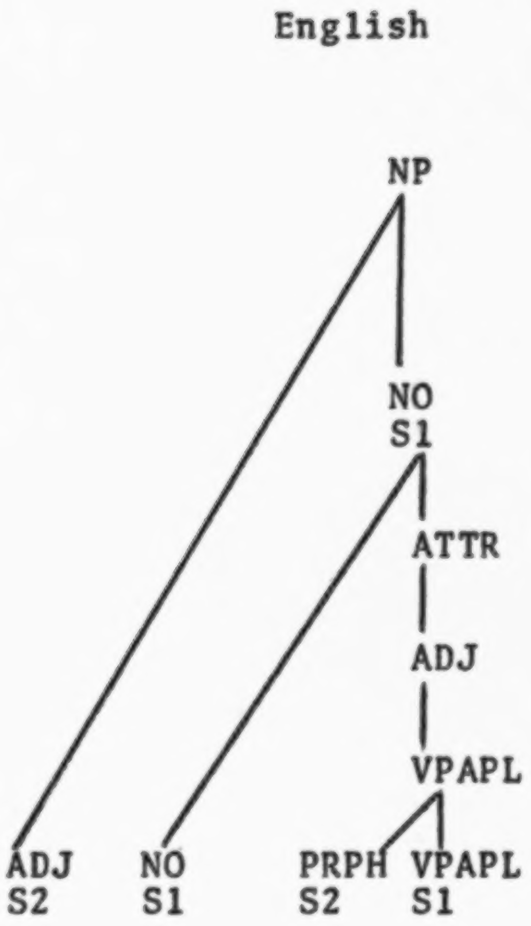
German *bestimmte durch die Kettenreaktion vorgeschriebene Dimensionen (ADJ + PRPH + PAPL + NO)*
English *definite dimensions necessitated by the chain reaction*

German *die von innenbuertigen Kraeften geschaffenen plumpen, stumpfen Formen*
English *the ungainly, blunt shapes created by interior forces*

In the first case (where the adjective precedes the prepositional phrase in German), English and German superscripts agree:


German

NP
ADJ
S2
ADJ
S1
A/X
VPAPL
ADJ
S2
PRPH
S2
VPAPL
S1
NO
S1

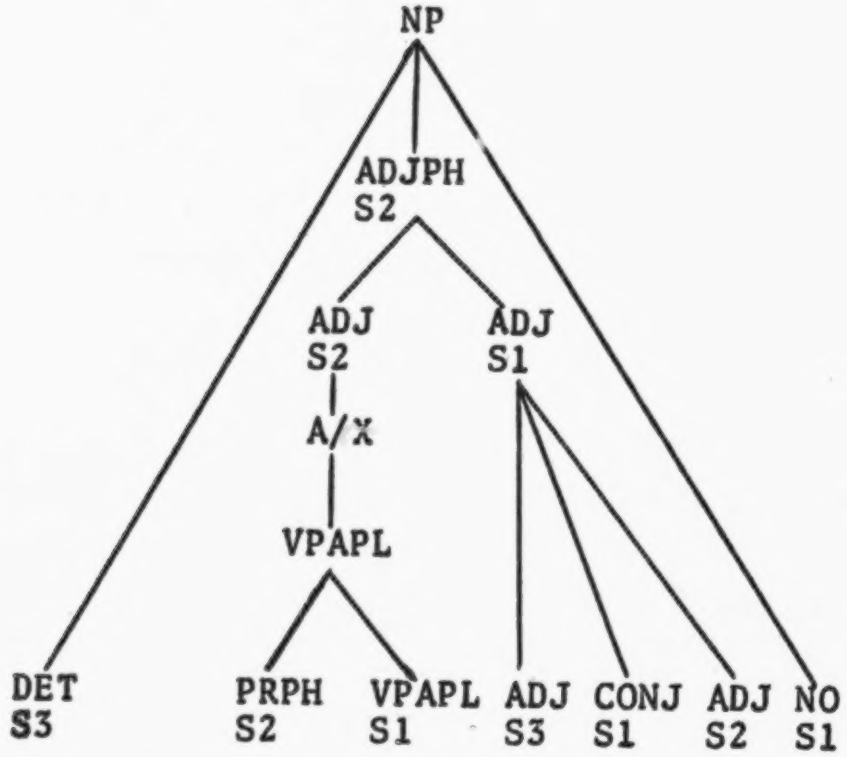

English

NP
NO
S1
ATTR
ADJ
VPAPL
ADJ
S2
NO
S1
PRPH
S2
VPAPL
S1

Derived Superscripts:
S4 S3 S2 S1 S4 S1 S2 S3

In the second case (where the adjective follows the prepositional phrase in German), superscripts do not agree and a multi-branch rule will be needed:

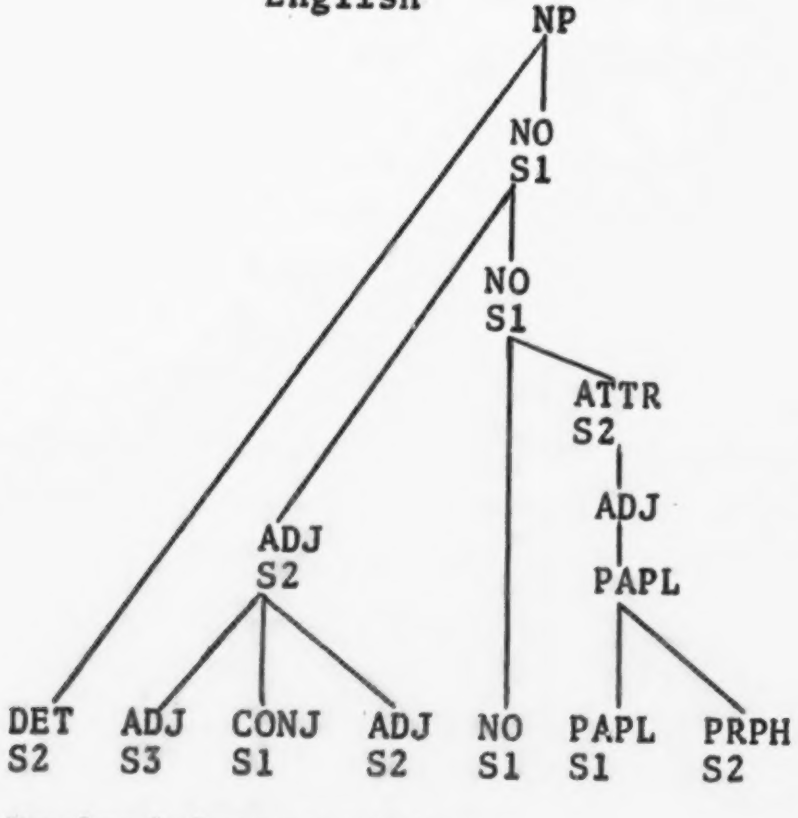
German


NP
ADJPH
S2
ADJ
S2
ADJ
S1
A/X
VPAPL
DET
S3
PRPH
S2
VPAPL
S1
ADJ
S3
CONJ
S1
ADJ
S2
NO
S1

Derived Superscripts:

S7 S6 S5 S4 S2 S3 S1

English


NP
NO
S1
NO
S1
ATTR
S2
ADJ
PAPL
ADJ
S2
DET
S2
ADJ
S3
CONJ
S1
ADJ
S2
NO
S1
PAPL
S1
PRPH
S2

Derived Superscripts:

S7 S6 S4 S5 S1 S2 S3

Past Participles Modified by Adverbs

These structures can frequently be translated into English pre-nominal or post-nominal structures. Examples:

German *aus einem derart konstruierten Reaktor*
English *from a thus constructed reactor*
or
from a reactor thus constructed

German *bei einheitlich zusammengesetzten Gesteinen*
English *in uniformly composed rocks*
or
in rocks uniformly composed

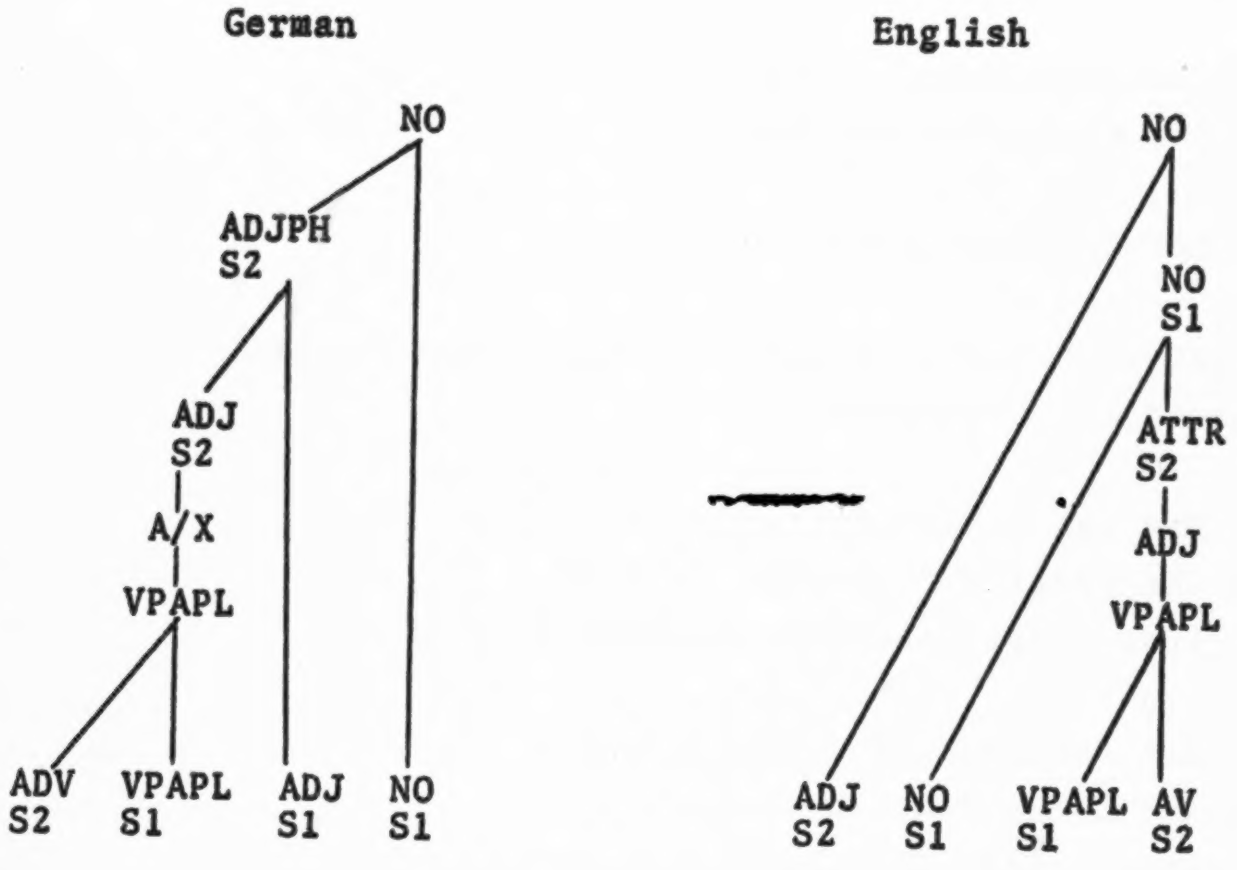
Other structures of the same type can be translated only into a post-nominal construction. Examples:

German *mit seewaerts gerichteten Bewegungen*
English *with movements directed seawards*
(not: seawards directed movements)

This seems to depend on the verb and adverb, i.e. dictionary items. After searching the available English corpora, it was decided that there are no such structures which must necessarily be translated into English pre-nominal constructions. We therefore wrote English syntax (or transfer) only for post-nominal structures of this type to avoid synthesis of *seawards directed movements*, *born later people*, *etc.*

As above (past participles modified by prepositional phrases), a multi-branch rule will be needed for occurrences of AV + PAPL and an adjective modifying a noun in order to guarantee correct superscripts. Example:

German *mit seewaerts gerichteten horizontalen Bewegungen*
English *with horizontal movements directed seawards*


German
NO
ADJPH
S2
ADJ
S2
A/X
VPAPL
ADV
S2
VPAPL
S1
ADJ
S1
NO
S1
English
NO
NO
S1
ATTR
S2
ADJ
VPAPL
ADJ
S2
NO
S1
VPAPL
S1
AV
S2

Derived superscripts (no agreement):

S4 S3 S2 S1 S4 S1 S2 S3

Unmodified Past Participles

Like the past participles modified by adverbs, the majority of unmodified past participles can be translated into pre-nominal or post-nominal English structures. Examples:

German *die ausgesendeten Alphateilchen*
English *the emitted alpha particles*
or
the alpha particles emitted

German *die erwaehnten Gasmassen*
English *the mentioned gas masses*
or
the gas masses mentioned

Others can be translated into an English pre-nominal structure only. Examples:

German *der verwoehnte Hoerer*
English *the spoiled listener*
(*not: the listener spoiled*)

German *die komplizierten Verbindungen*
English *the complicated compounds*
(*not: the compounds complicated*).

As in A. and B., multi-branch rules will be necessary if another adjective is involved and the past participle is to be translated into post-nominal position in English:

German *des ausgetriebenen reaktionsfaehigen Wasserstoffes*
English *of the reactive hydrogen expelled*

If in German the nominal is modified by a post-posed attribute (e.g. prepositional phrase or genitival construction) in addition to the pre-nominal past participle, the attribute usually follows the nominal directly in English and precedes the participle. Examples:

German *mit der ebenfalls freigewordenen Hydroxylgruppe der Base*
English *with the hydroxyl group of the base, also liberated*

German *die 1621 von Bachet de Mezirac veranstaltete Ausgabe des Diophant*
English *the edition of Diophant published in 1621 by Bachet de Mezirac.*

3.1.2.2 Post-nominal Attributes

A study of post-nominal attributes in the German physical science concordance excluded relative clauses and post-nominal infinitive constructions whose internal structure was not yet described. The following general patterns were found:

1.	NP-HEAD* + Genitive	450 occurrences
2.	NP-HEAD + PRPH	250 occurrences
3.	NP-HEAD + Appositive	190 occurrences
4.	NP-HEAD + Adjective	12 occurrences
5.	NP-HEAD + *usw.*	6 occurrences
6.	NP-HEAD + Adverb	2 occurrences

*NP-HEAD may be nominal, adjective phrase or demonstrative pronoun.

(1) Genitival constructions as post-nominal attributes:

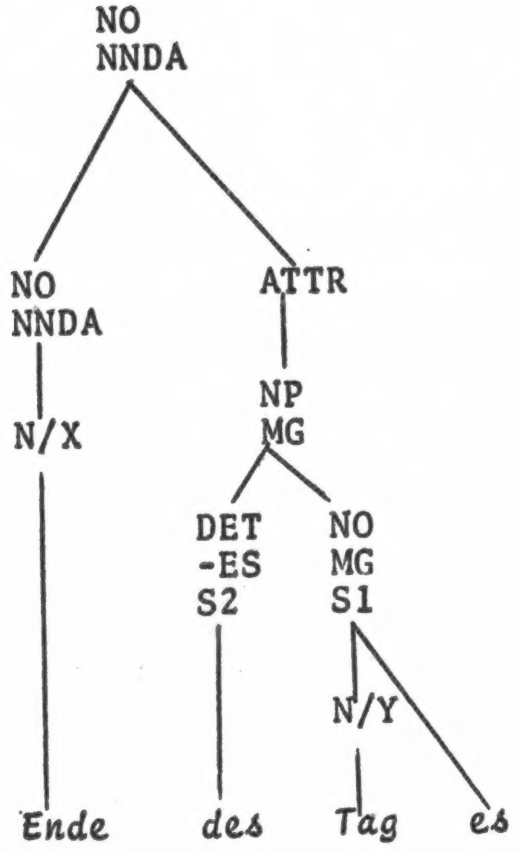
das Verhaeltnis der Groessen der Hindernisse

den Molekuelen der Luft

Strahlen robuster, langwelliger Konstitution

Ende des Tages

This pattern has already been described as


NO
NNDA
NO
NNDA
ATTR
N/X
NP
MG
DET
-ES
S2
NO
MG
S1
N/Y
Ende
des
Tag
es

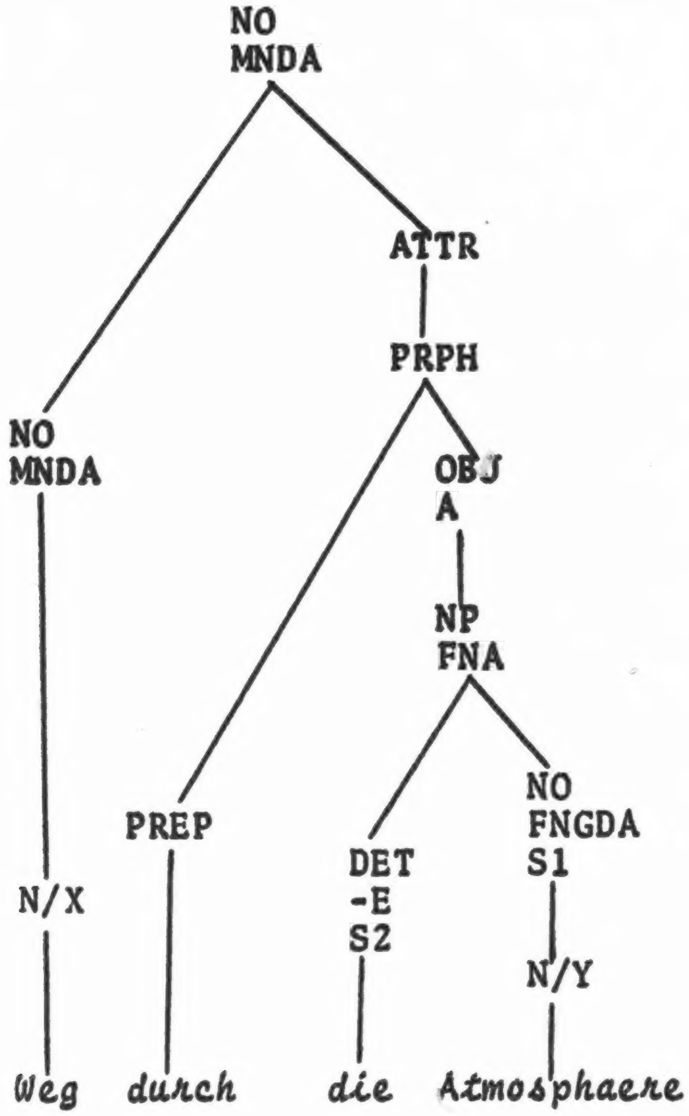
(2) Prepositional phrases:

der Weg durch die Atmosphaere

Annahme eines beliebigen Punktes
(innerhalb eines vorgegebenen Gebietes)

zehn Faecher dort - mit je zwei Knoepfen-

Analysis for the first (and most frequent) case had already been developed as shown:


NO
MNDA
ATTR
PRPH
NO
MNDA
OBJ
A
NP
FNA
PREP
NO
FNGDA
S1
DET
-E
S2
N/X
N/Y
Weg
durch
die
Atmosphaere

For description of the other examples given above, the following rules were developed:

```
VATTR  *(     VPRPH  *)
              B      B
VATTR  *-     VPRPH  *-
```

Post-nominal adjectives (including past and present participles) are usually set off by commas or sometimes dashes and can be preceded or followed by modifying adverbials, usually prepositional phrases.

Examples:

es wird neben dem Blauen auch langwelligeres Licht, gruen, gelb, rot gestreut

die Ausbreitungsgeschwindigkeit, unabhaengig vom ...

die Kohlensaeure - aus Wasserstoff, Kohlenstoff und Sauerstoff bestehend -

These examples were analyzed as post-nominal attributes, as described further below.

The items *usw.*, *etc.*, *u.dgl.*, *und dergleichen* were subsumed under the common label VETC, which was concatenated with SUBJ, OBJ, APP and ADJPH.

Examples:

die Grundbegriffe Gerade, Ebene usw.

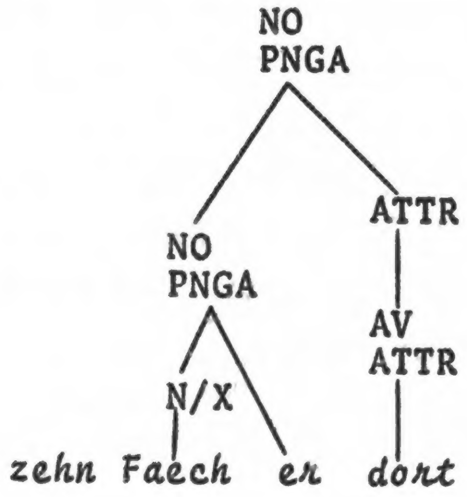
rot, gruen, gelb usw.

Post-nominal adverbs were already described in our grammar. Since not all adverbs can occur in this position, only a subclass of adverbs, AV/ATTR, is concatenated with the preceding noun phrase head.

Examples:

das Paar dort drueben

zehn Faecher dort


NO
PNGA
NO
PNGA
ATTR
AV
ATTR
N/X
zehn Faech er dort

Rules for the analysis of post-nominal adjectives were coded.

Examples:

langwelliges Licht, gruen, gelb rot

Ausbreitungsgeschwindigkeit, unabhaengig vom

The label VATTR was changed to VATTR/B in all rules of the form

NO NO ATTR

and then ATTR/B was defined as follows:

```
VATTR/B*       VATTR
               B

VATTR/B*,      VADJPH (*,)
                -0

VATTR/B*,      VADJPH VPRPH (*,)
                -0
```

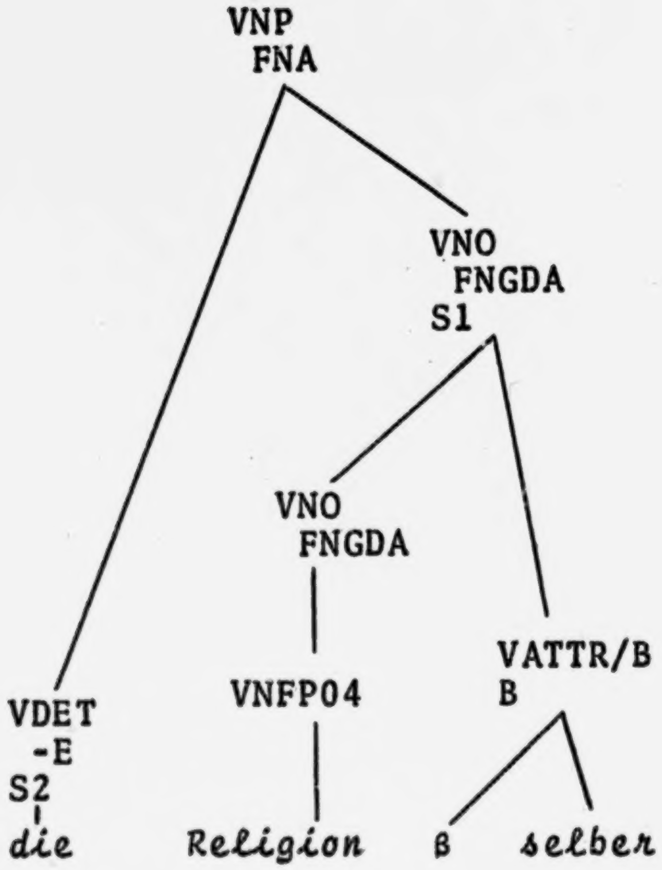
Since the number of NO-labels in German increased to 151, the above set of 5 rules saved several hundred rules which otherwise would have been necessary.

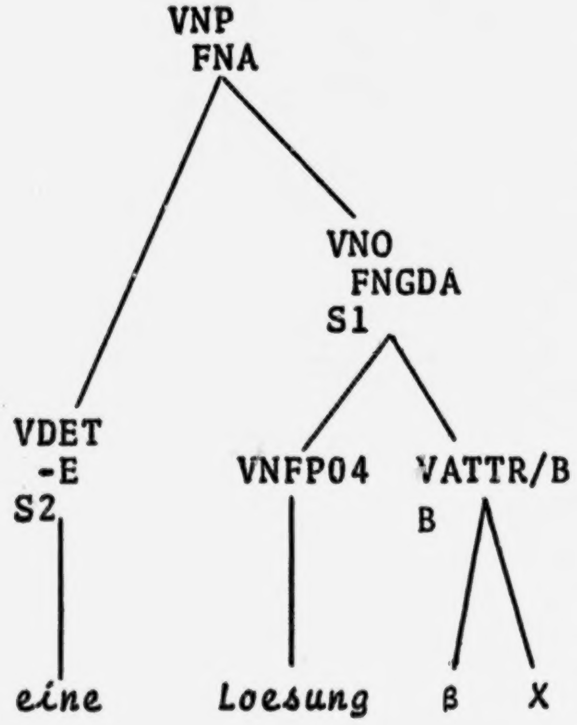
Also included in the class ATTR/B are the items *selber* and *selbst*, and letters of the alphabet, for analysis of structures like

die Religion selber,

eine Loesung X ... etc.

These items are entered in the lexicon with a leading blank to prevent separate analysis of every letter in every word.


VNP
FNA
VNO
FNGDA
S1
VNO
FNGDA
VNFP04
VATTR/B
B
VDET
-E
S2
die
Religion
ß
selber


VNP
FNA
VNO
FNGDA
S1
VDET
-E
S2
VNFP04
VATTR/B
B
eine
Loesung
ß
X

The symbol ß denotes the leading blank coded in the rules. All examples are taken from the German physical science concordance.

Analogously to the rules of the form

```
VATTR/B*,     VADJPH (*,)
               -0     B
```

The set of rules of the form

```
VATTR/B*,     VVPAPL (*,)
                X     B

VATTR/B*,     VVPRPL (*,)
                      B
```

were coded for analysis of post-nominal participles.

Example:

der Mann, die Hand ausgestreckt,

We coded these rules for all past participle labels in order to include participles modified by preceding adverbs or objects.

Two rules were coded analyzing participles as members of the adjective class AI/-+-. Members of this adjective class do not occur in uninflected form and, therefore, cannot be interpreted as complement. This prevents double analysis of structures such as

er ist gestuerzt

as VB + CMPL in addition to the correct verb phrase analysis.

GRM rules were coded for analysis of the uninflected, postnominally used adjective *voll(er)* which is usually followed by a genitive, less frequently by an accusative object:

eine Kiste voll(er) zerbrochener Flaschen

ein Atemzug voll Tabakrauch

These rules have the form

```
          VATTR   VVOLL   VOBJ
                           G or A
```

where VVOLL is defined as *VOLL and *VOLLER.

The grammar was coded for the concatenation of nominals with relative clauses in one rule:

```
          VATTR/B*,     VCLS
                         REL
```

This was possible because the relative clause label does not carry gender information.

3.1.2.3 Appositives

Because of case agreement, appositional constructions were not subsumed under the label ATTR which does not carry case information. The new label APP/CASE was created. It dominates the following structures which were found in appositional position:

Nominals, e.g.

unter dem Begriff apparative Strahlantriebe

die physikalische Masseinheit Elektronenvolt

der Begriff Dutzend

Noun Phrases

die Lehre von Raum und Zeit, die Kinematik,

auf einer Kurve, der "Abstandslinie"

Names (which are subsumed under nominal labels in our grammar)

der Mailaender Herzog Sforza

den Planeten Mars

Numbers and Letters

das Atom-Gewicht 222

August 1960

mit dem Betrag R

dem Argument Phi

Since there is no case agreement in these latter constructions, numbers and letters will be subsumed under the label ATTR.

The above are examples of appositives directly following the modified noun. Other appositional structures contain an item conjoining the noun and the following appositive.

Examples:

ein Nichtmetalloxyd wie Kohlenstoff

die Richtantenne bzw. das Interferometer

einer bestimmten Wellenlaenge bzw. Frequenz

About 1,000 rules concatenating nominals with appositions were encoded. Two modifications were made:

(A) Gender and number were not indicated in APP-symbols, since gender or number agreement is unnecessary for such constructions in German, e.g.

wohnten die Roemer, ein tapferes Volk

or

wohnten ein tapferes Volk, die Roemer

Subscripts of APPS are N, G, D, A, NG, NA, GD, GA, NGA, NDA, GDA, and NGDA (thirteen in all). This reduces the number of necessary rules by 75 per cent without loss in analysis or synthesis power.

(B) Nine concatenation types were listed:

VNO VNO VAPP
Gender Gender
Case Case

VCONJ/ VAPP
APP

VAV VCONJ/ VAPP
APP

Where the appositional construction can be preceded by comma or colon, or be preceded or followed by comma or dash. These were reduced to one type:

```
VNO       VNO       VAPP/B
 Gender    Gender    Case
 Case      Case     B
```

where APP/B directly dominates the above types. Encoding blanks as the initial right-side term allowed this solution. In such cases the third term has to have a blank operator, e.g.

```
VAPP/B *      VXYZ
 Case         B
```

Rule saving is about 60 per cent.

Rules were coded for the analysis of structures like

der Begriff dutzend,

die physikalische Masseinheit Elektronenvolt ...etc.

These structures have the form

```
VNO       VNO      VAPP/B
 Gender    Gender  .N(G)(D)(A)
 Case      Case
```

where

```
VAPP/B        *  VAPP
 N(G)(D)(A)       N(G)(D)(A)
```

and

```
VAPP/B        *, VCONJ/ VAPP
 N(G)(D)(A)       APP    NA
```

As in the example

> *mit ihren Ausdrucksmitteln, wie das kirchliche Bekenntnis*

Numbers

(a) Arabic Numbers -- There are several conventions for writing multi-digit Arabic numbers in German: 12795000, 12.795.000, or 12 795 000. The following set of rules was coded for analysis:

```
VDIGIT *1
VDIGIT *2
VDIGIT *3
......
VDIGIT *0

VNU    VDIGIT VDIGIT          Example: 12
                B
VNU-H  VDIGIT VNU                      121
                B
VNU-T4 VDIGIT VNU-H                    1212 or 1 212
               (B)
VNU-T5 VNU    VNU-H                    12121 or 12 121
               (B)
VNU-T6 VNU-H  VNU-H                  121212 or 121 212
               (B)
VNU-M7 VDIGIT VNU-T6               1212121 or 1 212 121
               (B)
```

etc.

Where NU-H stands for number hundred, NU-T4 for number thousand, four digits, NU-T5 for number thousand, five digits, etc. The distinction between the labels NU-T4, NU-T5, etc. was made to prevent the incorrect analysis and translation of, e.g. 12643 as NU-M which would be a result of the more general rules:

```
          VNU-T  VDIGIT VNU-H
                        B

and       VNU-M  VDIGIT VNU-T
                        B
```

(VNU-T would dominate 4 to 6 digit numbers)

The six rules above were written once with and once without blank operator, as indicated, and once with *. between the two right-hand members for analysis of all possible writing conventions for numbers with more than three digits.

For analysis of decimal numbers the following rules were written

```
          VDIGIT VDIGIT *,   VDIGIT        e.g.  1,2
                        B    B
          VDIGIT VDIGIT *,   VNU                 1,21
                        B    B
          VDIGIT VDIGIT *,   VNU-H               1,212
                        B    B
etc. to
          VNU-T6 VNU-T6 *,   VNU-T6              121212,121212
                        B    B
```

This set of rules was duplicated with */ in the place of *, for analysis of fractions.

Superscripts in all rules for analysis of Arabic numbers are 2 - 1.

(b) Spelled-out Numbers -- Rules necessary for analysis of spelled-out numbers up to the hundred thousands were coded. Again, superscripts are 2 - 1 with one exception - the rule for analysis of spelled-out two digit numbers:

```
VNU    VDIGIT *UND   VNU
 WR            B      WR
                     B
```

since in German spelled-out numbers the units precede the tens.

34 = *vierunddreissig*

In the German physical science concordance there were spelled-out numbers up to three-digit numbers and, of course, even hundreds, thousands, etc.

zweihundertfuenfzig

hunderttausend etc.

(c) Numbers Within Noun Phrase -- For analysis of strings like (DET) + NUMBER (+ ADJPH) + NO it was decided to concatenate NUMBER and ADJPH. The only difference in distribution between the combination NUMBER (+ ADJPH) and the simple adjective phrase lies in the fact that the former is followed by plural nouns only. Since this is a feature of synthesis, NUMBER and ADJPH were concatenated recursively under the label ADJPH. For the string

```
(DET) + NUMBER + NO
```

rules of the form

```
VADJPH VNUMBER
 -X
```

were coded.

Of all spelled-out numerals, only *zwei* and *drei* can take the genitive inflection *-er*. There is no occurrence in any of our corpora of the form *dreier*, but *zweier* occurs quite frequently. Adjectives following this form can be weakly or strongly inflected [2].

zweier unbestimmt gelassener Zahlen

zweier unbestimmten

For this reason it was classified as a DET and, to account for strong inflection in the adjective, it was treated as a constant concatenated with the plural genitive adjective:

```
VADJPH * ZWEIERVADJPH
 -ER            -ER
```

Grammar rules for the analysis of uninflected prenominal modifiers (*viel, wenig, etwas, nichts, so etwas, gar nichts* etc.) were coded:

VNP VADJ VNO
NI

etwas Penicillin

VADJPH VADJ VADJPH
NI

nichts Gutes

This second set was written recursively under the ADJPH node to provide analysis of the sequence ADJ/NI + A + N (*etwas frisches Penicillin*) as well as for ADJ/NI + A, which would otherwise require another (large) set of rules.

For structures in which these uninflected items occur alone, rules analyzing them as SUBJ/S and OBJ(D or A) were coded.

3.1.2.4 Conjunctions

Only three classes of conjunctions are necessary for analysis of German noun phrases.

(a) CONJ/APP which introduces certain appositives

(b) CONJ/1 the first item of discontinuous conjunctions

(c) CONJ/B Under this classname, we collapsed CONJ/A, CONJ/N and CONJ/2. Thus, no distinction between correlative and disjunctive conjunctions or between adjective-connecting and noun-connecting conjunctions is made since these features are unnecessary for analysis.

All items in the class CONJ/B have a leading blank or comma, e.g.

```
CONJ/B * UND
       *,
       *, ABER AUCH
```

Syntactic rules concatenating CONJ/B with other items were coded:

```
VSUBJ  VSUBJ  VCONJ/BVSUBJ
 S or P S or PB       S or P
VOBJ   VOBJ   VCONJ/BVOBJ
 CASE X CASE XB       CASE X
VCMPL  VCMPL  VCONJ/BVCMPL
              B
VAPP   VAPP   VCONJ/BVAPP
 X      X     B
```

Since the terminal conjunction classes are to be collapsed by macro-request, adverb and adjective concatenation rules, which had been coded earlier, also contain the new label CONJ/B. Superscripts in all the above rules are 3-1-2.

Duplicate analysis of two nominals joined by a comma, once as SUBJ or OBJ + CONJ/B + SUBJ or OBJ and once as NO + APP, cannot be avoided.

Etc., usw. u.dgl.

Since the above items occur in the corpus following single nouns as well as strings like NO + CONJ + NO, we decided to write two-branch rules (instead of one set of 4-branch and one set of 2-branch rules):

```
VSUBJ  VSUBJ  VETC
VOBJ   VOBJ   VETC
VAPP   VAPP   VETC
VADJPH VADJPH VETC
```

Superscripts are 2 - 1, analogous to those of CONJ rules. The terminal class ETC. contains the members *ETC., *USW, *U.S.W., *U.DGL., *UND DGL.

In order to avoid multiple analyses we decided to concatenate adjectives by multi-branch, non-recursive rules

```
VADJPH VADJ    VCONJ/BVADJ
 Inflec Inflec B       Inflec
 tion   tion           tion
        S3     S1     S2

VADJPH VADJ    VCONJ/BVADJ   VCONJ/BVADJ
 Inflec Inflec B       InflecB       Inflec
 tion   tion           tion          tion
        S5     S1     S4      S2     S3
```

The class CONJ/B includes comma. The above sets of rules were written with and without connecting conjunctions. The TRN labels are VCN+AJ+/AJ (D3) and V,+AJ+/AJ (D2). The longest string found in the German physical science concordance consisted of three adjectives.

The concordance was searched for strings of determiners. Only the two conjunctions *und* and *oder* were found to join determiners. Example:

dieser oder jener Strahl

Rules for analysis of determiner strings were, therefore, designed and coded as follows:

```
VDET     VDET      *UND   VDET
 Inflec   Inflec           Inflec
 tion     tion             tion
         S2               S1

VDET     VDET      *ODER  VDET
 Inflec   Inflec           Inflec
 tion     tion             tion
         S2               S1
```

Transfer was coded for the following constructions:

> Two adverbs joined by a conjunction. The grammar for this construction consists of a three-branch rule

```
VAV     VAV     VCONJ/BVAV
        S3      B      S2
                S1
```

This rule was placed in the new transfer class VCN+AV+ (D3).
AV

> Two appositional structures joined by a conjunction.

Examples:

die Zustandsgroessen Leuchtkraft und Temperatur

die beiden leichtesten Elemente Wasserstoff und Helium.

Since the nominals or noun phrases which constitute the appositive are not concatenated with the conjunction on a nominal or noun phrase level but rather on the subject-object level, a separate set of concatenation rules for examples like the ones given above had to be written. These rules have the form

```
VAPP   VAPP   VCONJ/BVAPP
       S3     B      S2
              S1
```

and were included in transfer under the label VCN+AP+$_{AP}$ (D3).

Several possible descriptions of noun compounds in German (On NO- and terminal level) were designed for occurrences like

die Sonne- und Mars-beobachtungen

Argon-Atome

Etc.

To facilitate the transfer coding for this kind of structure, we adopted that description which most closely resembles the English description.

Grammar rules were coded for description of object-participle constructions like

die kohleverbrauchende Industrie.

These rules have the form

```
VVINF   VOBJ   VVINF
        S2     B
               S1
```

and will thus also serve to analyze infinitives of this form, such as *das Kohleverbrauchen*. The already existent rule

```
VVPRPL  VVINF  *D
               B
```

will be applied to analyze the full present participle form as in the first example given above.

Grammar was also designed and coded for numbers in combination with *mal*, *plus*, *minus*, etc.

```
VNUMBER*PLUS  VNUMBER
VNUMBER*MINUS VNUMBER
VATTR  *DIVIDIVOBJ
        ERT DU A
        RCH
VATTR  *MULTIPVOBJ
        LIZIER D
        T MIT
```

This description was chosen to insure translation of *durch* as *by* rather than *through* in this environment.

Examples:

oberhalb minus 73 Grad C

plus 60 Grad C

acht mal sieben schritte

Tidenhub dividiert durch Wassertiefe.

3.1.3 Clause

Much of the effort put into German clause description is reported in the section, Problems in Grammar Design, (3.1.1). The following section reports other details of German clause structure.

Two different structural descriptions had been given for the environment of classnames dominating noun phrases or clauses, e.g.

der Herr sah es (SUBJ + PRED + OBJ)

and

der Herr sah, dass es gut war (SUBJ + PRED +*, + OBJ/CLS)
B

These can now be reduced to SUBJ + PRED + OBJ
B

where

VOBJ *BLANK VNOUN PHRASE
B

and

VOBJ *, VOBJ/CLS

This saves some multiple of clause rules, since subordinate-clauses can occur in any clause position and more often than once. To permit this simplification, blank-operators had to be added to the existing clause-rules. At the same time indication of agreement between syntactic classnames was dropped.

German clauses had previously indicated four types of agreement:

(1) Person agreement between subject and predicate

(2) Occurrence of proper auxiliary for past-participles

(3) Co-occurrence of modal and infinitive

(4) Object government by verbs (though this was implied via type (2) for past-participles, and not for finite verb and infinitive, since the necessary sub-classification for the German verb-dictionary with respect to types (2) and (4) does not yet exist).

As to (1) each clause rule so far exists in two versions (for some patterns three), 3rd person singular and 3rd person plural. The third was for first P.SG. A fourth will be necessary for 2nd P.SG which occurs, however, only seven times in the biological sciences corpora. First P.PL is subsumed under 3rd P.PL, while **2nd** P.PL does not occur, e.g.

VCLS	VADV	VVAX	VSUBJ	VVPAPL
		1ST	S	1ST
		S		

and

```
VCLS   VADV   VVAX   VSUBJ  VVPAPL
               1ST    S      1ST
               P
```

A count of the prefix *ab* in the physical sciences corpora showed that 75% of the occurrences were in prefix-verb sequences.

The distinction 'S, P, 1S' is now dropped from the clause level, which reduces the number of clause rules to 33 per cent or 25 per cent, if 2nd P.SG. is also counted. The agreement is provided through the transfer, by adding the steps

```
(A) VSUBJ VSUBJ
            S
```

and

```
VVB   VVB
       S
```

or

```
(B) VSUBJ VSUBJ
            P
```

and

```
VVB   VVB
       P
```

to the clause TRN-rules.

The number of TRN-rules thus remains the same. Some saving in encoding does, however, also occur, since a part of the TRN-rules remains constant.

As to (2) and (3) almost all clause rules contain a finite + A nonfinite verb-element on clause level. Each of these patterns exists in three versions for active (and passive):

(a) Co-occurrence of past-participle
+ auxiliary *sein*

(b) Co-occurrence of past-participle
+ auxiliary *haben*

(c) Co-occurrence of infinitive
+ modal

For passive

(a) Co-occurrence of past-participle
+ auxiliary *werden*

(b) Co-occurrence of *worden*
+ auxiliary *sein*

(c) As (c) above

Rules for patterns like *sein* + *zu* + infinitive (passive)
haben + *zu* + infinitive (active)
so far not written, would again have increased the number of clause rules.

Examples:

Die Linguisten haben einige Probleme zu loesen.

Er hat zu arbeiten.

Die Probleme sind zu loesen.

Sie sind nicht ganz einfach zu loesen.

This co-occurrence indication was also dropped, thus reducing the number of rules again by 33, or 25 per cent (with the last examples counted).

Clause rules now read like this

```
          VCLS    VADVRB VVFINITVSUBJCTVVERBAL
                   B-OP   B-OP   B-OP   B-OP
                           B      B      B
```

VFINIT → BLANK + the three auxiliaries and BLANK + MODAL (where
B-OP
the latter show person indication).
VERBAL → BLANK + past-participle, or + infinitive, or + 'zu' + inf.
B-OP

Naturally, encoding of these rules will take some time.

As to type (4) since the relative superscript of an object changes with its case (OBJECT/AKK always precedes OBJECT/DAT in superscripting), case had to be indicated on clause level.

The overall saving is thus a reduction to one from 12 (or 16 or 20, if 2nd person singular and plural are counted). The number of patterns is equal to the variations of type (1) times variations of types (2 + 3), or maximally of 95 per cent in clause rules.

With agreement thus dropped from clause level a totally tagmemic listing of adverbials, objects and complements on clause level (as originally attempted by Theo Vennemann [3]) is now feasible since it would only slightly increase the number of clause rules. This is probably the best solution. Since, however, domination of adverbials, complements and objects by infinitives and past participle is necessary for their adjectival and attributive use, (e.g. *das von den Roemern Zerstoerte Carthago* or *seine Absicht, fortan der Kunst sein Leben zu widmen*), patterns with initial subject were not coded tagmemically.

Rules for clauses containing up to three adverbs and two objects were encoded. Because of the different superscript assignation for subject in passive sentences (in presence of actor-adverb or an additional object) the subscript 'PASSIVE' was attached to the actor-adverb and the verb-elements in a passive sentence, since this set of rules had to be written anyhow. This means a duplication for the terminal auxiliaries and modals, but improves interpretation for all other verb elements.

Each pattern occurs in two variations depending on the position of the predicate:

(a) With final predicate

(b) With non-final predicate.

(b) again has three sub-types

(1) Finite predicate

(2) Finite predicate and prefix

(3) Compound predicate (AUX or MODAL + INFINITIVE)

All classnames carry the information B-OP on clause level, e.g.

```
SUBJCT , OBJECT , OBJECT , OBJECT , ADVERB , ADVERB
B-OP     B-OP     B-OP     B-OP     B-OP
A        D        G                 PASSIV
```

Each of them dominates a BLANK plus the identical symbol without 'B-OP'. Subject person indication is added, e.g.

```
VSUBJCT*     VSUBJCT
 B-OP          S
              B
```

```
Also   SUBJECT and OBJECT dominate
       B-OP        B-OP
                   A

       *,     VCLS
               DASS

       *,     VCLS   *,
               DASS  B

       VCLS   *,
        DASS  B
```

where

```
VCLS    VDASS   VCLS
 DASS            SUBORD
        S2      S1
```

and

```
VDASS   *DASS
        *OB
```

Similarly

```
VADVRB  *     VCLS   *,
 B-OP          SUBORDB
              B
        *,    VCLS
               SUBORD
        *,    VCLS   *,
               SUBORDB
```

The initial blank is provided at a rather high level, which only prevents illegal interpretations as clauses.

Rules were written for main and subordinate clauses. The corresponding passive structures were encoded by changing SUBJ to ADV/PASSIV, OBJ/A to SUBJ, and giving the verbal elements the subscript "PASSIVE" instead of "ACTIVE."

```
VCLS    VOBJECTVVFINITVSUBJCTVOBJECTVVERBAL
        B-OP    B-OP   B-OP   B-OP    B-OP
                ACTIVEB       D       ACTIVE
               B             B        -A
                                      B
        S2     S1     S5     S4       S2
```

becomes

```
VCLS    VSUBJCTVVFINITVADVRB VOBJECTVVERBAL
         B-OP   B-OP   B-OP   B-OP   B-OP
                PASSIV PASSIV D      PASSIV
               B      B      B       -A
                                    B
        S3     S1     S5     S4     S2
```

From these patterns passive structures without actor (ADV/PASSIV) were created by erasing 'ADV/PASSIV', e.g.

```
VCLS    VSUBJCTVVFINIT      VOBJECTVVERBAL
         B-OP   B-OP         B-OP   B-OP
                PASSIV       D      PASSIV
               B            B       -A
        S3     S1           S4     B
                                   S2
```

The grammar compilation program disregards the empty field, so that the rule thus is processed to the master store as

```
VCLS    VSUBJCTVVFINITVOBJECTVVERBAL
         B-OP   B-OP   B-OP   B-OP
        S3      PASSIV D      PASSIV
               B      B       -A
               S1     S4     B
                             S2
```

Structures for verb plus separated prefix were encoded. There is no corresponding passive structure.

There are two types of appositional sentences, those that begin with a conjunction and those that begin with a relative or interrogative pronoun, e.g.

Die Frage, ob er da sei.

Die Frage, warum er da sei.

Die Frage, die gestellt wurde.

The latter two have the same structure as VCLS.

3.1.3.1 Verb Phrase Rules

Two kinds of rules were written:

(a) Rules that connect clause level verbals which carry no subscripts as PERSON/NUMBER with those that do, e.g.

```
VVPRDCT*      VVPRDCT
 B-OP          ACTIVE
 ACTIVE        S
              B
```

(b) Rules that connect verbal with PERSON/NUMBER indication with verbals that indicate expansion possibilities, e.g.

```
VVPRDCTVVPRDCT
 ACTIVE ACTIVE
 S      EX
       S
```

This had to be done to reduce the number of TRN-rules for the time being because a rule that contains two variable elements X and Y, will have to be encoded 'X times Y' times. If the sequence is split into two rules, the number of necessary rules is X + Y.

Similarly verb-rules for verbals indicating agreement with auxiliaries or modals were encoded.

If in the future the longer strings (X times Y) are written to enable permutation transformations, which are possible only for TRN-rules with the same degree, the two-blob systems can be eliminated. The generalizing step in (a) and (b) can then be taken out by changing the 2nd term in type (a) to type (b), e.g.

```
VVPRDCT*       VVPRDCT
 V-OP           ACTIVE
 ACTIVE         EX
                S
               B
```

Type (b) was eliminated

Finally, the verb-phrase rules where verbals dominate adverbs and or objects were written. All of these are non-recursive, e.g.

```
VVPRDCTVADV   VVB
 ACTIVE         S
 FINAL S2     S1
 +ADV
 S
```

Verb classnames used on clause level are

VERBAL , B-OP	VERBAL , B-OP PASSIV	VPRDCT , B-OP	VPRDCT , B-OP PASSIV
VPRDCT , B-OP FINAL	VPRDCT , B-OP PASSIV FINAL	VFINIT , B-OP	VFINIT B-OP PASSIV

where the last two refer to auxiliaries and modals. The three non-passiv classnames (except VFINIT) may have the additional subscripts '-D' and '-A' which indicate permissible internal structure of the verb-phrase. This is necessary to guarantee correct superscript assignation for

Gestern hat er meinen Freund gesehen.

(as 3-1-5-4-2, instead of the incorrect 4-1-5-(3-2), which would occur if the expansion possibilities were not restricted dependent on the environment of the verb.

Thus for each basic verbal type four distinctions are made:

VVB , PERSON	VVPAPL,	VVINF,	VVINF ZU

(1) The verb is not expanded

VVB VVPAPL Etc.

(2) The verb is expanded by one or more adverbs

VVB
+ADV

(3) The verb is expanded by an accusative object

```
VVB   VVPAPL
 -D    -D
```

(4) The verb is expanded by a dative object

```
VVB   VVPAPL
 EX    EX
```

These are subsumed under the generalizing classes:

```
VPAPL  , VINF   , VINF   , VB
active   ACTIVE   ZU       ACTIVE
                  ACTIVE   S
```

and the corresponding passive classnames. This is necessary to avoid quadrupled encoding to account for the four types. However, a fully tagmemic clause level representation can always be simulated by transfer, eliminating this extra step.

The classnames subscripted 'PASSIV' dominate the first two types, as do those 'ACTIVE/-A'. 'ACTIVE' dominates all four, 'ACTIVE-D' the first three types.

The clause level verbal classes thus dominate

```
VVERBAL*    VVPAPL
             ACTIVE
            B

      *     VVINF
             ACTIVE
```

```
       *    VVINF
             ZU
             ACTIVE
            B

VVERBAL*    VVPAPL
 B-OP        -A
 -A         B

       *    VVINF
             -A
            B

       *    VVINF
             ZU
             -A
            B
```

There is no distinction ACTIVE necessary here.

```
VVERBAL*    VVPAPL
 B-OP        ACTIVE
 -D          -D
            B
       *    VVINF
             ACTIVE
             -D
            B

       *    VVINF
             ACTIVE
             ZU
             -D
            B
```

The other verbal classes are treated analogously. For analysis purposes $\text{VVINF}_{\text{SUBSCR}}$ and $\text{VVINF}_{\substack{\text{SUBSCR}\\ \text{ZU}}}$ could be collapsed. The distinction would, however, be necessary for synthesis of German.

3.1.3.2 Prefixed Verbs

For most verb classes, one transfer rule will suffice for a given prefix. The number is increased to two, however, if the verb class is dominated by *vinf*, because *zu* may be inserted, or by a participle with initial *ge*, e.g. *vor-trag(en)*, *vor-zu-trag(en)*, *vor-ge-trag(en)*. As can be seen from this example, the number of entries for a class will be three, if the verb class is dominated by both.

If the English translation has synonyms, the number of entries for a given prefix-verb combination will be multiplied by the number of these synonyms, since the TRN-classname of a lexical entry consists of a combination of the target- and source language canonical forms, e.g.

RECITE
=
VORTRA
GEN

and

REPORT
=
VORTRA
GEN

Thus also all the allomorphs of a verb, which necessarily belong to different verb classes in our system, will be listed under the same TRN-classname (or TRN-classnames for synonymous equivalents), e.g.

RECITE
=
VORTRA
GEN

contains

vor-trag, vor-ge-trag, vor-zu-trag,

vor-traeg, vor-trug, vor-trueg

as do all the other synonymous TRN-classes.

We estimate the number of prefixed verbs between 10,000 and 15,000. This number multiplied by some factor, dependent on above considerations, will be the number of TRN-rules for this particular problem.

To most efficiently make use of our macro-requests for linguistic data maintenance, we propose the following procedure to be executed under subsequent support.

The TRN-rule for P-V's consists of the rule-sequences:

(a)	VVERB	(VPRFX+VVERB)	(VVERB =*trag*)	(VPRFX=*vor*)
	CLASS	CLASS	CLASS	
	X1	X1	X1	

```
        (b)   VVPAPL (VPRFX*GE VVERB )  (VVERB =trag)  (VPRFX=vor)
               GE               CLASS    CLASS
                                X1       X1

or      (c)   VVINF  (VPRFX*ZU VVERB )  (VVERB =trag)  (VPRFX=vor)
               ZU-END          CLASS     CLASS
               PRFXT           X1        X1
```

The right sides of (a), (b), and (c) are identical, except for *GE in (b) and *ZU in (c). The necessary TRN-rules will consist of three syntactic rule numbers, of which only the first is different. This first number will be unique for each verb class, since it is dependent on it. We therefore need encode only one of the types (a), (b), (c). The other two types will be automatically created by REF,RT macro-requests, which copy the already established rules, but substitute the first member on the right side by the first member of the two other types. (Note: *zu* has to be picked up as a terminal entry, or else the degrees will not be the same. This, consequently, implies that the English *to* is picked up as a terminal, too. To avoid double encoding of English infinitives, *to* should be included in the next higher or any higher TRN-subtree.

A rule number (sequence) pertaining to the lexicon will be encoded as many times as it has translation equivalents. It suffices, however, to use a six-letter classname for the original encoding. The correct synonymous classnames can again be provided by REF,RT and one REF,CT macro. This is of considerable saving to the morphemes with many allomorphs, of which there is a minimum of two for each verb because of the *zu*-infinitive. It can also be encoded more easily. The six-letter classname uniquely represents each prefix-verb combination of a given class; the first two letters denote one of the 75 prefixes, the last four, the verb. The list of prefixes was obtained from *The Compound Verbs of the German Language* by Emmy Bauer [4].

3.2 Lexicography

Lexicographic effort during the contract was limited largely to compiling only those new entries needed for related syntactic testing. Casual maintenance was carried out on those existing entries found to contain errors.

Work was begun classifying all lexical items which occur in the German physical science concordance and are missing from our German dictionary. To date all such items from A to Q have been classified and entered into our system, including abbreviated forms, such as *Sek.* for *Sekunde*.

All letters of the Roman and Greek alphabets were entered into the dictionary as ATTR/B for analysis of letters in appositional position, e.g.

die Loesung X

These items were entered with a leading blank to prevent separate analysis of every letter occurring word-internally or word-finally. The written numbers *zwei* through *zwoelf*, as well as the forms *hundert* and *tausend* were classified and entered into the dictionary. All other written numbers are analyzed by morphological rules.

Terminal transfer was coded for possessive adjectives, conjunctions and all pronouns (demonstrative, personal, relative, interrogative, reflexive and expletive). The labels used for these transfer rules consist of the German and English canonical forms of the item being classified. For example, the membership of the transfer class

```
VER
 =
 HE
DO
```

consists of the forms

er, *sein*, *ihm*, *ihn* etc.

Approximately 2200 terminal transfer rules were coded for adjectives which in German have the prefix *un*.

Examples:

```
VUNANGEC67    C1234
 NEHM
 =
 UNPLEA
 SANT
DO
```

In this case the underlying monolingual rules are

```
C67    VAI/O+S*UN    VAI/O+S
                     B

C1234  VAI/O+S*ANGENE
                HM
```

Work was begun on transfer coding for uninterrupted prefix-verb sequences:

```
       VANHEUECR1    CR2    CR3
        RN
        =
        HIRE
       DO
```

Where the underlying rules are defined as:

```
CR1    VPAPL  VPRFX  *GE    VV27A
        GE-    B-OP  B      B
              S2            S1

CR2    VV27A  *HEUER

CR3    VPRFX  *AN
        B-OP
```

the inflectional ending *-t* of the full form *angeheuert* will be added by the next-higher rule

```
       VPAPL  VPAPL  *T
               GE-   B
```

German-English transfer rules for the German concordance of physical science corpora were systematically worked out to entry letters *sto*, checking in every case with existing entries to avoid duplication. Additional aids used include the physical science corpus, the monolingual lexicon, the zero transfer display, and various reference works [13, 14, 15, 16, 17].

Rules entered include:

(a) One-word correspondences or phrases which function as a single, nonvariable unit.

(b) The past tenses of strong verbs.

(c) The various meanings of a single word, insofar as these different meanings actually appear in the concordance data.

(d) Some proper names of persons, cities, topographical areas, etc., which can be expected to occur with reasonable frequency (e.g. *Muenchen* = Munich).

The entries exclude:

(a) All pronouns (demonstrative, personal, relative, interrogative, reflexive, expletive).

(b) Possessive adjectives

(c) Prepositions

(d) Modal verbs

(e) Separable verbs

(f) Reflexive verbs

(g) Prefixes and suffixes

(h) Unusual proper nouns.

Also excluded are:

(i) Words which have no exact one-word or nonvariable-unit correspondences.

(j) Words which regularly translate differently according to context (e.g., *beliebig*).

(k) Specialized, though not necessarily infrequent, meanings of words where inclusion of the specialized meanings would result in a false translation in the great majority of cases, e.g. *erschien = published, nehmen = get*, as in the phrase
Woher nahm Kepler die Laterne.

(l) Words (in phrases) of which the literal or figurative meanings would result in nonsensical or grotesque translations, e.g.

Ein Atemzug voll Rauch wird ... geblasen.

(m) Words or phrases for which the best transfer rules have not been decided upon, e.g.

los sein,

Schritt fuer Schritt.

Indications have simultaneously been made in the concordance as to which words have been entered either now or previously and which (i.e. those categories listed immediately above) require further work before being entered. Finally, some corrections of errors (typographical and/or lexical) in the zero transfer and the monolingual lexicon have either been made or indicated.

3.3 Test Corpora

Two types of concordances were displayed for use in linguistic description in German: the usual concordance, which shows each lexical item of a corpus (with its environment) in alphabetic order and reverse concordances, which display each lexical item in alphabetic order beginning with the last symbol of each item. Normal concordances available total approximately 248,300 words, reverse concordances total 87,300 words.

Corpora totalling approximately 521,300 running words were compiled. The corpora for which concordances were prepared deal with physical, biological and social sciences, more specifically with physics, mathematics, geography, geology, geophysics, astronomy, astrophysics, chemistry, nuclear medicine, cybernetics, psychology and sociology, economics and political sciences.

A glossary with frequency count was prepared for a total of 57,300 words.

4 ENGLISH

English design completed during the contract period in all major areas is presented in the following sections.

4.1 Syntax

4.1.1 Noun Phrase

Noun phrase description is essentially complete for the major areas discussed below.

4.1.1.1 Determiner Strings

Revisions in the rules concatenating strings of determiners were made. Previously such strings were treated under recursive binary rules. Currently they are treated under n-ary rules to preserve the ordering inherent in the terminal classification. The validity of this treatment has been borne out by comparison with treatments found elsewhere, [5,6,7]. There are now eight basic order classes, a class for the article *a*, a class for *an*, a class of pre-determiners, and a class of pre-articles.

Numbers were originally considered to be determiners, but a re-examination of their distribution demonstrated that they occur as pre- and post-nominal modifiers, as do adjectives. Therefore, numbers have been removed from the determiner classes and subsumed under the label for adjective. Rules to cover various other number patterns have also been written.

A large number of transfer classes involving determiners have been mapped and the rules coded as has transfer for numerals.

4.1.1.2 Adjective Strings With Connectors

This problem area includes all pre-nominal adjectives connected by commas and/or conjunctions. Such adjective strings are analyzed recursively under the label adjective. To facilitate analysis, a new conjunction label has been introduced: CONJ/B. The "B" indicates that a B-operator (blank suppressor) should be used. CONJ/B contains members such as:

```
*,
*,  VCONJ
*   VCONJ
```

Therefore, adjectives with connectors can be analyzed by concatenating adjective plus CONJ/B plus adjective.

Transfer covering these grammar rules has been written under the label CN+AJ+AJ.

4.1.1.3 Adjective Strings Without Connectors

Adjective strings without connectors are generated recursively from the noun head. Participles acting as modifiers in pre-nominal position are classed as adjective. Transfer over these adjective concatenation rules falls into the class NO+AJ.

4.1.1.4 Post-nominal Modifiers

Post-nominal modifiers are modifiers which occur immediately after the noun head. They include the following:

(a) Attributes -- there are now six types of modifiers under the label attribute: infinitives, adjectives, prepositional phrases, adverbs, present participles, and past participles. Attributes are analyzed by right-hand recursive rules from the noun head. Post-nominal modifiers are concatenated with the noun before pre-nominal modifiers are added.

Transfer related to the above rules is coded under the labels NO+AT, AV, and PRPH.

(b) Appositives -- appositives are distinguished from attributes by having commas after the noun head. There are two types of appositional structures presently handled by the grammar: those with a preceding adverb and those without. Appositives are joined to the nominal in the same manner as attributes. Transfer for appositives is labeled NO+AT.

(c) Relative Clauses -- relative clauses are a third kind of post-nominal modifier. They are concatenated like attributes and appositives.

4.1.2 Verb Phrase Description

Research on verb phrase level constructions was begun in October. Search of the scientific corpora yielded the patterns to be accounted for. For all occurring permutations of the elements, auxiliary, main verb, adverb, object, it was decided to concatenate all occurrences of auxiliary and main verb, then to concatenate with adverb, and finally to concatenate with object.

4.1.3 Clause

Rules to concatenate verbal complements with a form of the verb *be* were written, as well as rules expanding complements into all nominals, noun phrases, and adjectives. Rules to generate present and past tense distinctions were removed. It was decided that these tense distinctions would be accounted for in transfer, thereby allowing the syntactic description to remain relatively simple.

From the scientific corpora 100 sample sentences were chosen at random and clause patterns were outlined. On the basis of this analysis, plus information derived from secondary sources [8, 9, 10, 11, 12], clause patterns for subordinate, relative, and non-yes/no interrogative clauses were discussed and coding was begun on all types. In addition rules incorporating distinctions between final and non-final sentence position were written for all relevant clause elements. This was done in an effort to guarantee proper concatenation of final and non-final punctuation.

Transfer rules for structures with clause, clause relative and clause interrogative were coded. In addition the inflectional level transfer rules for present and past forms of verbs were coded.

As a result of demonstration testing in February 1966 it was decided to make some major revisions in the English grammar. As an aid to determining what revisions were desirable, an experimental grammar based on a relatively small corpus was prepared. We concentrated on reducing the large number of multiple analyses produced during the demonstration. Preliminary changes already indicated as necessary were carried out in March and consisted of the following:

(a) Development of the present LRS lexicographic classes.

(b) Mnemonic changes on the syntactic level to eliminate or merge classes where possible and to change abbreviations to coincide with the new LRS mnemonics.

(c) Elimination of all unnecessary generalizing steps in the grammar, that is, all one-branch rules.

4.1.3.1 Interrogative Clauses

Research on interrogative constructions in scientific corpora was begun in August. Very few interrogatives occur in this type of discourse, but it was found that there is a certain type of declarative statement that assumes an interrogative construction embedded in a larger syntactic unit. These declaratives are more frequent than the interrogatives whose form they take. For this reason it was suggested that the structure be described without any distinction made between the interrogative and declarative phrases.

4.1.3.2 Subordinate Clauses

A survey of the lexical class of subordinate conjunctions was made to determine the relevant distinctions between members of this class and members of the class of simple conjunctions. An additional survey was made of the scientific corpora to determine external clause patterns for the concatenation of dependent and independent clauses. It was found that the existing grammar would handle the patterns encountered, and that the membership of the subordinate conjunction class, conjunctions which introduce subordinate clauses (e.g. *if*, *although*, *since*, *even if*, etc.) was distributionally dissimilar to that of the simple conjunction class (e.g. *and, but, or,* etc.), as was expected.

4.2 Word Formation

4.2.1 Lexicography

Multiple classes consisting of adjective, noun and verb classes were introduced. For example, *abuse* can be both a noun and a verb. Instead of classifying it twice, we now code it in one multiple class called

VNX
VX

In conjunction with this change, the mnemonics were revised for the sake of clarity as well as to incorporate additional linguistic information.

Nouns now follow the pattern NWWWXY, where WWW=0001-999, an arbitrary serial number, X=C (consonant onset), V (vowel onset), or N (article information not relevant). Y=H (human), N (non-human), or B (both human and non-human). Verb and adjective classes follow a similar pattern, VWWWX and AWWWX.

The 5000 nouns changed to the intermediate classification scheme will be converted under subsequent support to the present form by means of macro-requests. From March to August 17,200 additional entries were converted. In August two IBM 1050 teleregister units were installed, enabling the lexicographers to type entries directly into the system.

This process speeded up the output by three times, and from August to October 25,800 entries were converted, totalling 43,000 entries.

4.2.2 Webster Morphology

Webster morphology is the set of inflectional rules resulting from our earlier system of lexicographic classification. Entries in our dictionary, which use the previous system of classification, are based on Webster's Collegiate Dictionary. The gaps which existed in both grammar and transfer have been filled and these rules are now complete.

As a result of the development of a new scheme of classification used in the RMD/LRS conversion, it became necessary to code a second set of inflectional rules for the grammar. For example, *snow* may be classed as adjective, noun, and verb, thus

AX
NX
VX

Since the number of such taxonomically complex classes theoretically possible is quite large, it was decided to code inflectional rules on classes defined by the lexicographers on the basis of utility and not on the basis of all possible combinations of adjective, noun, and verb features.

Currently, approximately 3500 morphological rules and an equal number of transfer rules are being maintained over 500 RMD/LRS terminal classes. The morphological rules supply:

(a) Singular and plural inflections over classes denoted by NWWWXY, where X=either an indication of consonant or vowel onset or an indication that article information is not relevant, and Y=either an indication of human or non-human or an indication that both distinctions are relevant.

(b) Present singular, present plural, past singular, past plural, present participle, and past participle inflections over classes denoted by VWWWX, where X=the same as in (a) above.

(c) Positive, comparative, superlative, and adverb inflections over classes denoted by AWWWX, where X=the same as in (a) above.

Transfer classes over the above inflections are, respectively:

(a) SG and PL
(b) PR (includes both singular and plural),
PA (includes both singular and plural),
-ing, and -ed
(c) PSTV, CMPRTV, SPRLTV, -ly.

4.3 Test Corpora

Thirty-five English corpus displays are available including about one million running words. There are 21,150 samples with about five lines per sample and 10 words per line.

Areas of discourse include metallurgy, chemistry, physics, geology, zoology and such non-scientific areas as economics, political science, psychology, travel. Statistically, special emphasis has been given to physics in the topics of astronomy, geology, acoustics and light.

4.4 Concordances

There are twenty-two normal-sort concordances available. One corpus has a reverse concordance. In all there are about 300,600 key words. There are also three large glossaries available.

5 CONCLUSION

In the area of descriptive effort tests on the syntactic data yielded satisfactory results for the noun phrase. The data are comprehensive for most constructions excluding the relative clause. The work done so far on the verb phrase has yielded similar results, although much remains to be done in order to have a comprehensive description. While considerable work has been done on clause constructions, lower level descriptions such as noun and verb phrase must be completed before the clause can be tested adequately. The effort will be proposed for continued support.

Under the contract reported here, no support was requested or provided for programming systems development. Therefore, we have reported nothing pertaining to systems problems. During the contract period, however, it became clear that LTS operating on the current hardware configuration was too severely restricted to permit completely adequate operation with the large data bases we have already accumulated. Future operations are expected to continue unimpeded on the CDC 6600/1700 system, components of which were recently acquired by the Computation Center and Linguistics Research Center, respectively, of the University of Texas.

REFERENCES

1. Quarterly Progress Report, LRC 66 P-27, Austin: Linguistics Research Center, May 1966.

2. Duden Grammatik der deutschen Gegenwartssprache, Bibliographic Institute, Mannheim, 1959.

3. Theo Vennemann, "German Syntax," LRC 66 WD-2, Austin: Linguistics Research Center, January 1966.

4. Emmy Bauer, The Compound Verbs of the German Language, Julius Groos, Heidelberg, 1926.

5. Carlyle Westbrook Barritt, "The Order Classes of Modifiers in English," (unpublished Ph.D. dissertation), University of Virginia, 1952.

6. C. M. Millward, "Rules for the Co-occurrence of English Pre-adjectival Noun Modifiers," (unpublished Ph.D. dissertation), Brown University, 1952.

7. O. Thomas, Transformational Grammar and the Teacher of English, Holt, Rinehart and Winston, New York, 1965.

8. William M. Austin, "A Linear System of English Syntax," Seminar Work Paper MT-53, Georgetown University Project in Machine Translation, 1957.

9. F. W. Householder, "Syntactic Clauses Used Currently in the Determination of Clause and Phrase Boundaries in English," Bloomington: Indiana University.

10. Otto Jesperson, Essentials of English Grammar, University of Alabama Press, 1964.

11. Ralph B. Long, The Sentence and Its Parts, University of Chicago Press, 1961.

12. J. P. Thorne, M. Whitfield, and R. M. Griffiths, "A Phrase Structure Recognition Routine/An English Grammar," Computer Unit Report No. 3, Edinburgh: University of Edinburgh, April 1964.

13. The New Cassell's German Dictionary, Funk and Wagnalls, New York, 1958.

14. Karl Wildhagen and Will Héracourt, *German-English Dictionary*, Brandstetter Publishing Company, Wiesbaden, 1957.

15. *Duden Orthography of the German Language and Foreign Words*, edited by Paul Grebe, Bibliographic Institute, Mannheim, 1958.

16. F. A. Brockhaus, *Der Sprach-Brockhaus*, Wiesbaden, 1961.

17. Louis DeVries, *German-English Science Dictionary*, McGraw-Hill, New York, 1959.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Linguistics Research Center, The University of Texas Box 7247, University Station Austin, Texas 78712	**2a. REPORT SECURITY CLASSIFICATION** Unclassified **2b. GROUP** N/A

3. REPORT TITLE

Research in German-English Mechanical Translation

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Final, 1 January 1966 - 30 December 1966

5. AUTHOR(S) *(Last name, first name, initial)*

Lehmann, Dr. W. P., Tosh, Dr. L. W.

6. REPORT DATE April 1967	**7a. TOTAL NO. OF PAGES** 120	**7b. NO. OF REFS** 17
8a. CONTRACT OR GRANT NO. AF30(602)-3991 *b.* PROJECT NO. 4599 *c.* *d.*	**9a. ORIGINATOR'S REPORT NUMBER(S)** LRC 67 AFSC 4 **9b. OTHER REPORT NO(S)** *(Any other numbers that may be assigned this report)* RADC-TR-67-98	

10. AVAILABILITY/LIMITATION NOTICES

This document is subject to special export controls and each transmittal to foreign governments, foreign nationals or representatives thereto may be made only with prior approval of RADC (EMLI), GAFB, NY 13440.

11. SUPPLEMENTARY NOTES Computational linguistics in German-English MT R&D Zbigniew L. Pankowicz/Project Engineer	**12. SPONSORING MILITARY ACTIVITY** Rome Air Development Center (EMIIH) Griffiss Air Force Base NY 13440

13. ABSTRACT

The report presents results of 12-month R&D in German-English mechanical translation on syntactic level. The effort was aimed at expanding a syntactic translation capability within the framework of the Language Translation System (LTS). Additional transfer grammar rules, linking two monolingual (German and English) syntactic descriptions, were designed and coded. German and English grammar data compiled under previous contracts were expanded. Grammatical descriptions are in context-free phrase-structure form with transformational facility provided in the structure of the interlingual transfer coding. The grammars were verified in the processing system against samples of text data in German and English. Ten paragraphs of German text were translated on trial basis.

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Machine Translation R&D						
Computational Linguistics						
Syntax of Natural Languages						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8*a*. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
________________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
________________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
________________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, *or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.